INPUT REPRESENTATIONS MEET PRETRAINING OBJECTIVES: A CONTROLLED STUDY OF GENERATIVE EHR FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurately modeling irregular event timing and numerical measurements is central to zero-shot prediction with generative Electronic Health Record (EHR) foundation models. Recent work has studied the impact of input event representations and loss function separately, overlooking their correlation and joint impact. Additionally, there is no comprehensive benchmark that systematically evaluates zero-shot inference performance on a large and diverse number of tasks. To address these problems, we conduct a control study that aligns the tokenizer, backbone, and training budget to isolate the effect of input embeddings and pretraining loss across three model families: Discretized, Continuous-fused, and marked temporal point-process (MTPP). For zero-shot evaluation, we introduce `ZEHRO` [1], a comprehensive zero-shot benchmark that spans two datasets, three task types, and two time horizons. Our results show that continuous representations and loss functions benefit regression, while discretization leads in classification. The point-process approach achieves the best performance in predicting large time intervals and remains competitive in long-horizon time-to-event prediction. Our ablation studies further show that changing the pretraining objective generally has a larger effect on performance than changing the input embeddings, and that these changes can have aligned or opposing effects across different tasks. These findings reveal task-dependent trade-offs in input representations and objective choices, informing their joint design in future generative EHR foundation models.

1 INTRODUCTION

Inspired by advances in language modeling, recent work represents electronic health records (EHRs) as sequences of clinical event tokens and pretrains generative foundation models using next-token prediction (Renc et al., 2025; Zhang et al., 2025; Waxler et al., 2025). Applying this approach to EHRs requires addressing two central challenges: irregular event timing across multiple temporal scales, which conveys semantic information about patient health; and heterogeneous events that include categorical medical codes with numerical measurements, such as laboratory results. These features motivate specialized input representations and pretraining objectives for event timing and numerical values.

Existing approaches make different design choices for these components. **"`Discretized`"** approaches map prediction time intervals and numerical measurements to bins, trained using cross-entropy (Renc et al., 2025). **"`Continuous-fused`"** approaches encode numerical inputs directly and combine categorical code prediction with conditional density estimation for numerical targets (Loza et al., 2025; McCann et al., 2026). Marked temporal point-process (**`MTPP`**) formulations alternatively model event types and their occurrence times with intensity functions (Shmatko et al., 2025; Jing et al., 2026). These choices concern distinct aspects of the model: how inputs are represented to the model and how loss objectives are optimized.

However, there is a lack of rigorous controlled studies examining how input embeddings and loss functions affect zero-shot performance. Existing generative EHR foundation models differ in their

[1] Code is available at `https://anonymous.4open.science/r/ethos_fm-27FE/`.

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

pretraining datasets, tokenizers, backbones, training procedures, and evaluation tasks. These differences make it difficult to isolate the individual and joint effects of input representations and pretraining objectives.

In this work, we conduct a controlled study of input representations and pretraining objectives in generative EHR foundation models. We align the training data, tokenized sequence, transformer backbone, and training budget, and compare three configurations: Discretized, Continuous-fused, and marked temporal point process (MTPP). We also introduced a comprehensive zero-shot evaluation benchmark that spans different task types and time horizons. To understand how input embedding correlates with the pretraining loss, we further cross discrete and continuous numerical inputs with categorical and Gaussian numerical objectives in a 2×2 factorial study. All models are evaluated in zero-shot settings on classification, event-timing, and numerical regression tasks, using both one-step prediction and long-horizon simulation.

Our contributions can be summarized as follows:

• **Unified zero-shot benchmark.** We proposed **ZEHRO**, a comprehensive zero-shot evaluation benchmark for EHR foundation models. This benchmark spans three types of tasks: classification, time-to-event, and regression, with both one-step prediction and long-horizon simulation. We benchmark all models on MIMIC-IV and an institutional EHR dataset for external validation.

• **Controlled comparison.** We establish a controlled pipeline that uses the same tokenized sequence layout, model backbone, and training setup, and exclusively study the impact of input representations and prediction objectives.

• **Ablation Studies of input embeddings and pretraining loss.** We separate the effects of input encoding and numerical objectives through a 2×2 ablation study, showing that objective changes produce larger performance differences than input embeddings. We also demonstrate that the correlation between input embeddings and pretraining loss is task-dependent.

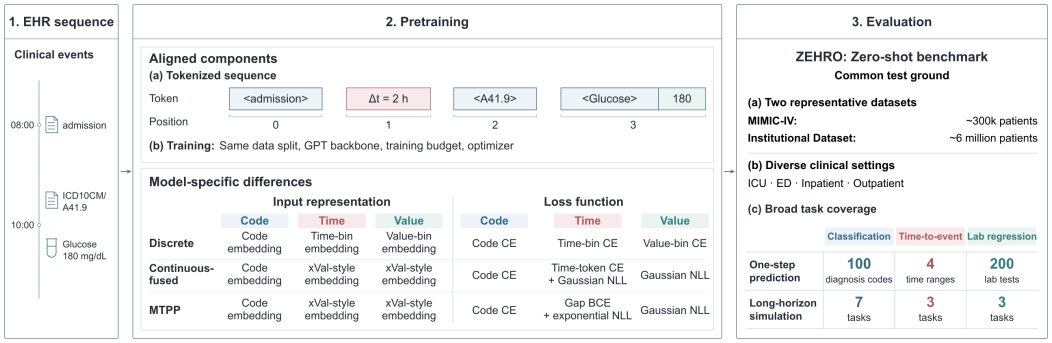


Figure 1: Overview of our analysis pipeline

2 Related Work

Generative EHR foundation models Generative foundation models adopt different input embeddings and pretraining loss functions. Both ETHOS (Renc et al., 2025) and Curiosity (Waxler et al., 2025) discretize time intervals and lab values into bins and use cross-entropy loss for next-token prediction. They are pretrained on MIMIC-IV and Epic Cosmos, respectively. To preserve the continuous information, multivariateGPT Loza et al. (2025) combines code and numerical value embeddings and uses a Gaussian negative log-likelihood loss for numerical prediction. Experiments on small clinical datasets show improved regression performance over discretized baselines. To handle the irregular time intervals, Delphi (Shmatko et al., 2025) applies homogeneous Poisson processes to jointly predict disease occurrence and timing. Trained on UK Biobank data, it demonstrates strong performance in disease risk and time-to-event prediction. However, differences in training datasets and downstream tasks limit direct comparisons across these methods (Table 1). A comprehensive benchmark under the same dataset is needed to systematically compare their input representations and pretraining objectives.

Input Embeddings of EHR Foundation Model EHR consists of mixed event types at irregular time intervals, including numerical and non-numeric codes. Discrete time tokens (Pang et al., 2021), continuous time embeddings (Kazemi et al., 2019), or attention-based rotary embeddings (Su et al., 2024) have been proposed to model the irregular time intervals or absolute time of events. For numerical values, discrete quantiles Renc et al. (2025) and continuous representations have been used to encode numerical values. Controlled comparisons indicate that their effects depend on the encoding and evaluation setting. Guo et al. (2026) found that attention-based rotary embeddings beat time tokens in most classification tasks. To encode values, Lee et al. (2026) discovered that discrete and soft-discretization encoders outperform xVal for both classification and regression tasks, while Montgomery et al. (2026) shows that continuous embeddings are more precise in arithmetic calculation. One key confounding factor in these studies is that most methods have varied lengths of tokenized outputs. Even with the same pretraining budget, the number of patients seen in the pretraining varies across methods. To isolate the impact of sequence length, we therefore introduce a shared tokenized sequence across methods in section 5.2.

Pretraining Loss of EHR Foundation Model Pretraining objectives specify the distributions that a model learns to estimate. When a single medical trajectory is represented by a vocabulary of discrete tokens, discrete-token approaches optimize categorical likelihoods through cross-entropy (Renc et al., 2025). To preserve the original numerical value, a few works directly model the continuous probability with either a Gaussian distribution (Loza et al., 2025) or a normalizing flow (Draxler et al., 2026). Event timing has also been modeled using survival and point-process objectives (Shmatko et al., 2025; Steinberg et al., 2024). Under a fixed tokenizer, Jing et al. (2026) report improved downstream performance with a marked time-to-event objective relative to next-token prediction on their tasks. However, this paper uses the same tokenizer across different losses. Our factorial study varies both the input embeddings and loss functions to examine the separate and joint effects of these choices.

Table 1: Comparison of Zero-shot Inference Tasks in Previous Literature. Mixed Time Horizon means whether the evaluation tasks include both one-step prediction and long-horizon simulation.

Model	Classification	Time-to-event	Regression	Mixed Time Horizon
ETHOS (Renc et al., 2025)	✓	✗	✗	✗
Curiosity (Waxler et al., 2025)	✓	✗	✗	✓
Delphi-2M (Shmatko et al., 2025)	✓	✓	✗	✓
MultivariateGPT(Loza et al., 2025)	✗	✓	✓	✗
ZEHRO(Ours)	✓	✓	✓	✓

3 Problem Formulation

For each patient $i \in [N]$, we represent the observed electronic health record (EHR) as a time-ordered sequence of clinical events:

$$\mathcal{H}_i = ((\Delta t_{i,j}, e_{i,j}))_{j=1}^{N_i}$$

where N_i is the number of observed events for patient i. $\Delta t_{i,j} = t_{i,j+1} - t_{i,j}$ equals to the elapsed time since last medical event. We assume that events are ordered chronologically, such that $\Delta t_{i,j} \geq 0$, allowing multiple events to share the same timestamp. Each event $e_{i,j} = \{m_{i,j}, v_{i,j}\}$ contains a clinical code $m_{i,j} \in \mathcal{M}$, and an optional numerical value $v_{i,j} \in \mathbb{R} \cup \{\varnothing\}$. The code $m_{i,j}$ may, for example, represent an ICD-10 diagnosis, a procedure, a medication, or a laboratory measurement. $v_{i,j} = \varnothing$ indicates that no numerical value is associated with the event.

We denote the event history preceding the j-th event by $\mathcal{H}_{i,<j} := ((t_{i,\ell}, e_{i,\ell}))_{\ell=1}^{j-1}$. A natural modeling objective is to maximize the joint likelihood of the observed event sequence:

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \prod_{j=1}^{N_i} p_\theta \left(\Delta t_{i,j}, m_{i,j}, v_{i,j}, |\mathcal{H}_{i,<j} \right) \tag{1}$$

4 ZEHRO: ZERO-SHOT EVALUATION BENCHMARK

We propose ZEHRO, a zero-shot benchmark that evaluates three components of an EHR trajectory: clinical codes (classification), event timing (time-to-event) and numerical lab tests (regression), as shown in Figure 2. All evaluations use frozen pretrained parameters without task-specific fitting. In one-step prediction, models predict the next target from the observed history. In long-horizon simulation, we follow (Waxler et al., 2025) to generate $N = 64$ trajectories in parallel and derive predictions from the simulated trajectories using task-specific stopping rules. We report AUPRC for classification tasks, normalized MAE for time-to-event tasks, and R^2 for regression tasks.

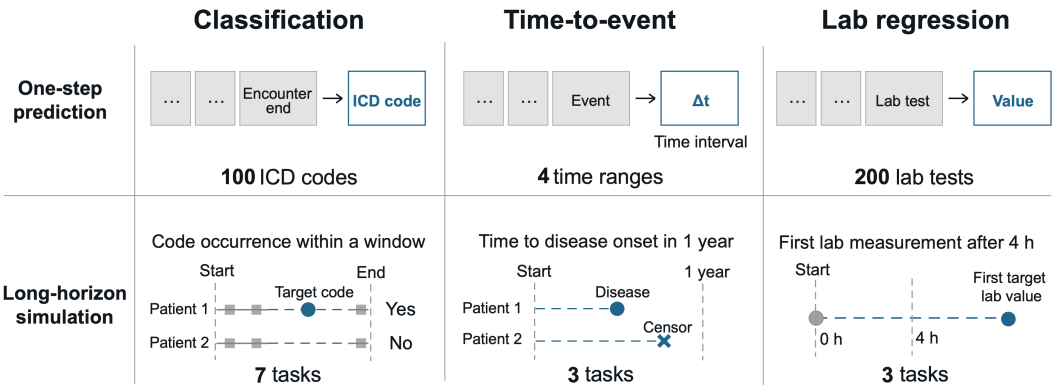


Figure 2: Visualization of tasks in ZEHRO Benchmark

4.1 ONE-STEP PREDICTION

Classification Given the observed history through the end of each encounter, we predict the encounter's primary diagnosis and report the mean one-vs-rest AUPRC across 100 diagnosis codes.

Time-to-event We predict the elapsed time to the next recorded event with a later timestamp, evaluating positive intervals across four ranges of observed duration: 1 min-1 hour, 1 hour-1 day, 1 day-1 month, 1 month-2 years.

Lab regression Given the observed history and target laboratory code, we predict the associated numerical value across 200 laboratory measurements.

4.2 LONG-HORIZON SIMULATION

Classification We design 7 binary classification tasks.

• 4 clinical outcome prediction: in-hospital mortality, prolonged hospital stay (> 10 days), critical outcomes within 12 hours of emergency department (ED) triage, and ED reattendance within 72 hours of discharge.

• 3 disease onset prediction within one year: acute myocardial infarction (AMI), metabolic dysfunction-associated steatotic liver disease (MASLD), and stroke.

Time-to-event Using the same disease-specific cohorts as classification, we evaluate the time to the first recorded diagnosis of AMI, MASLD, or stroke.

Lab regression We select three important lab tests related to sepsis patients: creatinine, glucose, and platelets. We randomly select a time during ICU stays that has no target lab measurement in 4 hours. Predictions are obtained from the first occurrence of the target test in each simulated trajectory.

5 CONTROLLED STUDY

Our benchmarking builds on three representative generative foundation models, from which we extract distinct representation and pretraining loss choices for a comprehensive evaluation: ETHOS (Renc et al., 2025) motivates the use of "Discretized" tokenization, MultivariateGPT (Loza et al., 2025; McCann et al., 2026), which encodes numerical data directly with categorical code prediction and numerical density estimation, which we refer to as "Continuous-fused", and Delphi-2M (Shmatko et al., 2025) uses an MTPP pretraining loss (called MTPP). One bottleneck for directly comparing these models is that each uses different datasets, tokenization, model backbones, dataloaders, and training hyperparameters. We align these stages to isolate the impact of input representations and pretraining loss on downstream zero-shot performance. Note that the original Delphi-2M paper excludes numerical measurements from the loss function and handles concurrent events by masking concurrent events to attention computation. We revise their input representations and pretraining losses to enable numerical event prediction, thereby improving its performance. We also remove their mask tie to use the same causal attention algorithms as others, which is necessary for autoregressive generation during zero-shot evaluation. Our detailed setup is as follows.

5.1 DATA

We pretrain different models from scratch on two EHRs: MIMIC-IV 2.2 Johnson et al. (2023) and an EHR dataset from a large urban hospital called 'Institution' hereafter. MIMIC-IV 2.2 contains records from approximately $300k$ patients treated at Beth Israel Deaconess Medical Center between 2008 and 2019. We include all available hospital, intensive care unit (ICU), and emergency department (ED) records. The Institution dataset contains 6.7 million patients between 1986 and 2023, including primary care, inpatient visits, and outpatient visits data, representing a more general EHR dataset than MIMIC-IV. During preprocessing, we clip all numerical value to 1st-99th percentile of the training data distribution, same as McCann et al. (2026). We clip positive time intervals to between one minute and two years for MIMIC-IV, reflecting its focus on acute hospital care. For Institution, which spans both inpatient and outpatient care, we use a broader range of one minute to 100 years to accommodate longer gaps between encounters. Additional details of Institution are in Appendix A.1.

5.2 INPUT REPRESENTATIONS

All three models shared identical event stream formats for tokenized sequences: $\mathcal{H}_i = ((\Delta t_{i,j}, e_{i,j}))_{j=1}^{N_i}$ for each patient i. The time interval token is inserted before an event whenever $\Delta t_{i,j} > 0$ and is omitted between concurrent events. Each clinical event $e_{i,j}$ occupies one position in the transformer: the clinical code and its optional numerical value are fused into one embedding. $\Delta t_{i,j}$ is omitted for concurrent events. Each event $e_{i,j} = \{m_{i,j}, v_{i,j}\}$ includes medical code type m and optional numerical mark v associated with the code. For each randomly sampled patient timeline, we follow Renc et al. (2025) to add three demographic tokens (sex, age at the start of the record, and race) at the start and one token `TIMELINE_END` at the end. All models have the identical position layout in the tokenized sequences. The key differences are how these models represent time intervals and numerical measurements.

Discrete representation We follow ETHOS (Renc et al., 2025) to map time intervals into $N_q = 20$ fixed time bins, and each of them is a unique time token in the vocabulary (e.g., 0-5 minutes, 5-15 minutes, etc.) Numerical measurements are assigned to $N_q = 20$ empirical quantile bins. For a code $m_{i,j}$ with value-bin Q_{v_k}, the event embedding is $E[e_{i,j}] = E[m_{i,j}] + E[Q_{v_k}]$, where $E[m_{i,j}]$ is the code embedding and $E[Q_{v_k}]$ is a shared bin embedding across different measurements. We also explore how different discretization choices $N_q = \{10, 20, 50\}$ impact the downstream performance and present our analysis in Appendix A.4.

Continuous representation We use the same continuous representations for continuous fused and MTPP methods. To keep the continuous value as inputs, we follow McCann et al. (2026) to do the normal-score transformation of each numerical value $z(v_{i,j})$. For any numerical event $(m_{i,j}, v_{i,j})$, the embedding is $E[e_{i,j}] = E[m_{i,j}] + \text{GeLU}(W_{m_{i,j}} * z(v_{i,j}) + b_{m_{i,j}})$. This is similar to the xVal method (Golkar et al., 2023), except that each numerical code has unique linear projection matrices $W_{m_{i,j}}$ and $b_{m_{i,j}}$ instead of sharing the same across codes. The time interval is embedded with the same continuous embedding method.

5.3 LOSS FUNCTION

Following the autoregressive decomposition of the likelihood in Equation 1, the loss function is:

$$\ell(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{N_i} \underbrace{-\log p_\theta(\Delta t_{i,j} \mid H_{i,<j})}_{\ell_{i,j}^{\text{time}}} \underbrace{-\log p_\theta(m_{i,j} \mid H_{i,<j}, \Delta t_{i,j})}_{\ell_{i,j}^{\text{code}}} \underbrace{-\log p_\theta(v_{i,j} \mid H_{i,<j}, \Delta t_{i,j}, m_{i,j})}_{\ell_{i,j}^{\text{value}}} \tag{2}$$

averaged over positions. The code term $\ell_{i,j}^{\text{code}}$ is a categorical cross-entropy over the medical codes in all models.

Discretization Similar to ETHOS (Renc et al., 2025), the discrete arm uses cross-entropy over all time bins and value quantiles. Denote $Q_{t_{i,j}}$ and $Q_{v_{i,j}}$ as the observed bin for observed time $\Delta_{t_{i,j}}$ and value $v_{i,j}$, the loss function is:

$$\ell_{i,j}^{\text{time}} = -\log p_\theta(Q_{t_{i,j}}), \qquad \ell_{i,j}^{\text{value}} = -\log p_\theta(Q_{v_{i,j}}). \tag{3}$$

Continuous-fused We follow multivariateGPT (McCann et al., 2026) to fit the normal-scored time $z(\Delta t_{i,j})$ and values $z(v_{i,j})$ with Gaussian distributions. For each distribution, we fit two trainable linear projection heads for mean and variance: $\mu(\cdot), \sigma(\cdot)$. The losses are:

$$\ell_{i,j}^{\text{time}} = -\log \mathcal{N}\big(z(\Delta t_{i,j}); \mu_t(H_{i,<j}), \sigma_t(H_{i,<j})\big) \tag{4}$$

$$\ell_{i,j}^{\text{value}} = -\log \mathcal{N}\big(z(v_{i,j})\big); \mu_t(H_{i,<j}, \Delta t_{i,j}, m_{i,j}), \sigma_t(H_{i,<j}, \Delta t_{i,j}, m_{i,j})\big) \tag{5}$$

The transformation $z(\cdot)$ is determined by the prior distribution selected from a few candidates $\{$normal, log-normal, gamma, Weibull and min-max$\}$. For any numerical observation, we fit it to each candidate distribution and evaluate the normal-score transformation. We select the candidate with the smallest QQ-score and use the distribution for pretraining stages. During inference, we sample a numerical value by inverse transformation $z^{-1}(v_k)$, where v_k is drawn from its fitted Gaussian distribution. To handle concurrent events, both the discretization and continuous fused approaches add time tokens to the next-token prediction vocabulary for medical codes. Time advances only when a time-interval token is sampled.

Marked Temporal Point Process (MTPP) Following Delphi-2M (Shmatko et al., 2025), we model the next event as the first arrival of competing homogeneous Poisson processes. Each code $m_{i,j}$ has a unique intensity function $\lambda_{m_{i,j}}$, with the total rate $\Lambda = \sum_{m_{i,j}} \lambda_{m_{i,j}}$. Different from previous loss functions, the type of next code $m_{i,j}$ is conditionally independent of the time interval $\Delta t_{i,j}$ given the history $H_{i,<j}$. The positive interval follows $\Delta t_{i,j} \sim \text{Exp}(\Lambda)$.

One limitation of the homogeneous Poisson is that it struggles to model concurrent events where $\Delta t_{i,j} = 0$. It masks out all concurrent events during attention computation, which excludes valid information like diagnoses or lab tests at the same timestamp. Instead, we introduce a Bernoulli gate $\rho(H_{i,<j})$ that first decides whether time has advanced at all, which allows us to keep the causal self-attention as in other models. The time loss function is the sum of the Bernoulli gate loss and the Exponential NLL of poisson process:

$$\ell_{i,j}^{\text{time}} = \text{BCE}\big(\rho(H_{i,<j}), \mathbb{1}[\Delta t_{i,j} > 0]\big) + \mathbb{1}[\Delta t_{i,j} > 0]\big(\Lambda\, \Delta t_{i,j} - \log \Lambda\big)$$

For the value loss function, we use the same Gaussian NLL as the one in the `Continuous-fused` approach.

5.4 TRAINING AND EVALUATION

All models use the same GPT-2-style decoder-only transformer (Radford et al., 2019) with a context window of 2048. Following the same backbone configuration as Renc et al. (2025), we use 6 layers, 12 attention heads, and set the model width to be 768. We train all models up to 30k steps and choose the checkpoint with the lowest validation loss for zero-shot evaluations. As all models shared the exact same layout for tokenized sequences, they are trained with an equal number of patient timelines. The detailed training and hyperparameter procedure can be found in Appendix A.2.

5.5 Factorial Study of Representations and Objectives

We conduct a 2×2 factorial study crossing numerical input encoding (discrete bin embeddings or continuous scalar embeddings) with the numerical prediction objective (categorical cross-entropy or Gaussian negative log-likelihood). This yields four configurations: Discrete-Categorical (same as "`Discretized`"), Continuous-Categorical, Discrete-Gaussian, and Continuous-Gaussian (same as "`Continuous-fused`"), where the first term denotes the input encoding and the second the numerical objective. We set "`Discretized`" as the baseline and calculate the performance gap between the other three ablations and it to study the impact of input embeddings and pretraining loss:

• **+Continuous Input Embedding:** This is the performance gap between Continuous-Categorical and the `Discretized` method, reflecting the impact of replacing discrete input embeddings with continuous ones.

• **+Continuous Fused Loss:** This is the performance gap between Discrete-Gaussian and the `Discretized` method, reflecting the impact of replacing cross-entropy with a Gaussian distribution.

• **+Continuous Input Embedding & Continuous Fused Loss** This is the performance gap between **Continuous-fused:** and the discrete method, reflecting the impact of jointly changing the input embeddings and loss functions to continuous versions.

6 Results

6.1 Benchmarking Results on MIMIC-IV and Institution Datasets

`Discretized` leads classification prediction, whereas `Continuous-fused` leads regression across datasets. We pretrain and evaluate each model separately on MIMIC-IV and the Institution dataset, and evaluate them on the `ZEHRO` benchmark. Figures 3 and 4 show consistent classification and regression trends across datasets. For classification, `Discretized` achieves the highest aggregate AUPRC at both horizons. `Continuous-fused` performs similarly in one-step prediction but falls behind in long-horizon simulation, while `MTPP` has the lowest aggregate scores. For regression, `Continuous-fused` achieves the highest aggregate R^2 at both horizons, followed by `MTPP` and `Discretized`. These results indicate a consistent task-dependent trade-off: discretization favors classification tasks, whereas continuous numerical modeling favors regression tasks in both horizons. The only difference between `MTPP` and `Continuous-fused` is their loss function for time. The results indicate that the point process hurt both the classification and regression performance.

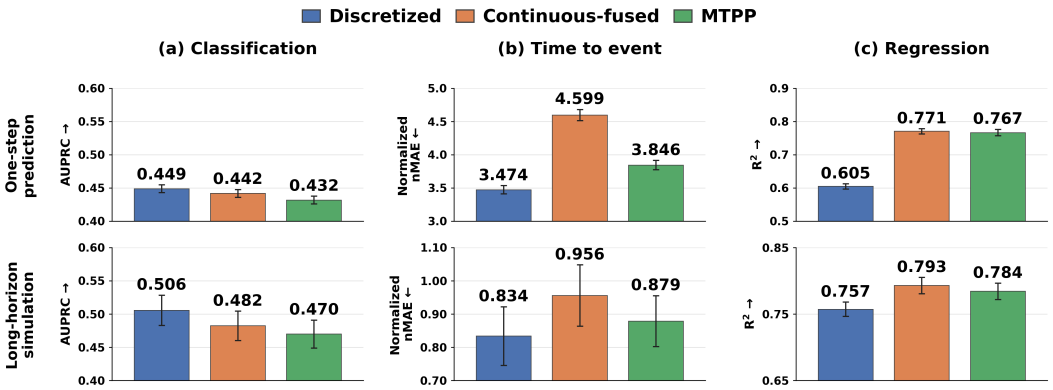


Figure 3: Comparison of Zero-shot Performance Between `Discretized`, `Continuous-fused` and `MTPP` method on MIMIC-IV. We report AUPRC for classification, normalized MAE for time-to-event, and R^2 for regression tasks.

Time-to-event performance varies across datasets and prediction horizons. On MIMIC-IV, `Discretized` achieves the lowest timing error at both horizons, followed by `MTPP`

and `Continuous-fused`. On the Institution dataset, the gap between `Discretized` and `Continuous-fused` narrows, particularly in one-step prediction. This can be explained by the differences in the empirical time interval distributions (see Figures 9 and 10 in Appendix A.6). The time interval distribution in MIMIC-IV exhibits pronounced peaks at 1 minute, 15 minutes, and 1 hour, which are highly correlated with the hospital's recording routine. This spiky distribution may favor the `Discretized` method, as it allows flexible allocation of probability across time bins to accommodate these irregularities, whereas a single continuous distribution cannot capture this pattern well. However, the Institution has a smoother and long-tailed distribution, making it easier to fit the Gaussian distribution in `Continuous-fused`. The `MTPP` approach has divergent performance across time horizons on Institution. It achieves the highest aggregate one-step prediction error but the lowest average long-horizon error. We stratify the interval into different time scales and calculate the normalized MAE of each group. The stratified results in Table 9 in Appendix A.5 show that `MTPP` performs worst in all bins below 30 days, particularly below one hour, but best for intervals between 30 days and two years. This pattern suggests that the point-process approach is better suited to predicting large time gaps than short ones, which also explains its competitive performance in long-horizon simulation tasks.

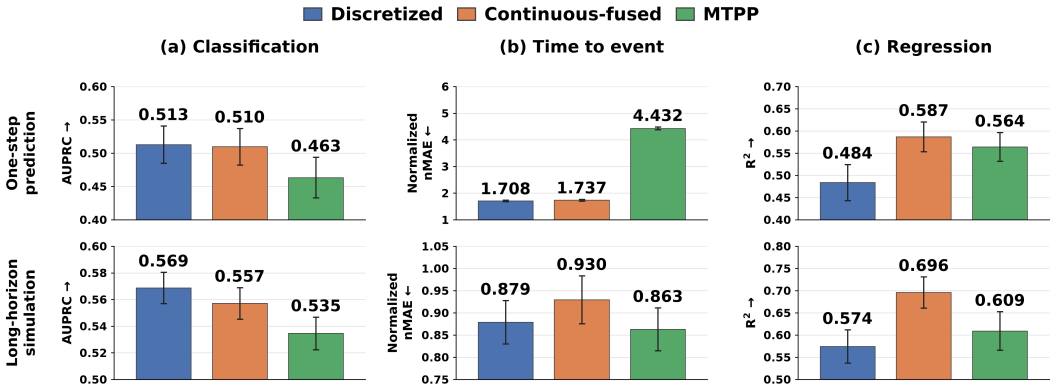


Figure 4: Comparison of Zero-shot Performance Between `Discretized`, `Continuous-fused` and `MTPP` method on Institution dataset

Test-time Compute At inference time, we scale the number of parallel trajectories N generated for Monte Carlo estimation of a single code or value and make plots in section A.3. We find that for all models the performance consistently improves as we scale up N and reaches a near plateau at $N = 32$ or $N = 64$, which is consistent with findings from previous literature (Waxler et al., 2025).

6.2 Ablation Study of Input Representation and Loss Function

The pretraining loss has a larger effect than the input representation. Starting from the fully discrete model, we replace either the input representations or the loss function with a continuous version, or change them at the same time. The performance in Figure 5 shows that using the continuous loss function has a much greater impact on performance than adding continuous input embeddings across all evaluation tasks, in both one-step and long-horizon simulations. For example, for one-step time-to-event prediction, changing the loss function shifts the performance by 1.237, which is 16 times the impact of changing input embeddings (0.076). Similarly, in regression tasks, the loss change has approximately seven times the effect of the input embeddings.

Input embeddings and pretraining losses have aligned or opposing effects across tasks. Relative to the fully discrete baseline, changing the input embeddings and pretraining losses to their continuous counterparts does not always affect performance in the same direction. For classification, continuous embeddings alone produce little change, whereas continuous losses reduce performance at both horizons. For regression, both changes improve performance estimates, and their combination achieves the highest R^2 at both horizons. In contrast, the continuous versions of input embeddings and loss functions have opposing effects on one-step event timing but aligned effects in long-horizon prediction. Adding the continuous loss function only hurts the classification and time-to-event

performance while benefiting the regression. This indicates a trade-off of learning different task signals during optimization. Meanwhile, we notice the incompatibility issue between the input embedding and the loss function. If we fix the loss function type, choosing the same input embedding type benefits long-horizon tasks, where the generated events are encoded as input for future rollouts. For example, the `MTPP` model always outperforms the ablation with the same continuous loss but discrete input embeddings. For the `Discretized` model, replacing it with the continuous input embedding hurts time-to-event performance while having minimal impact on classification and regression tasks.

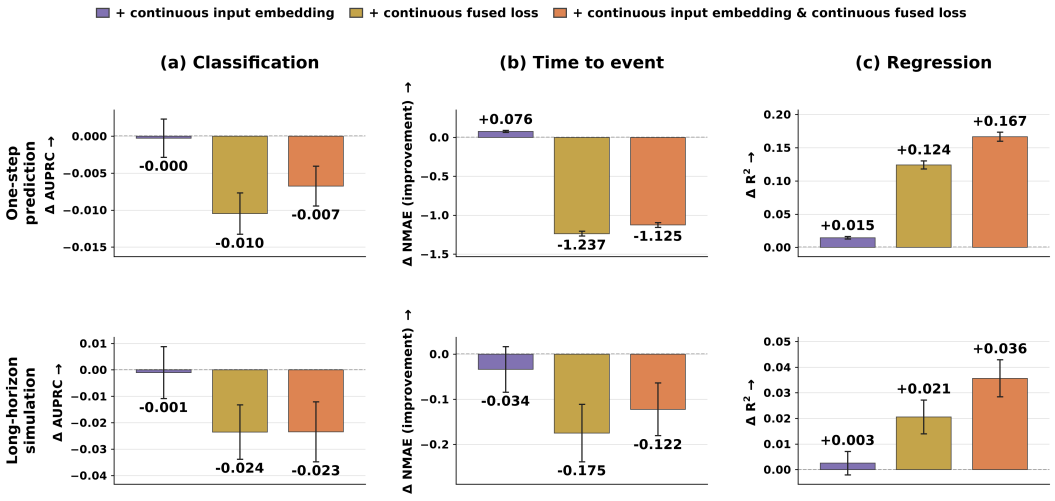


Figure 5: Zero-shot Performance of Ablation Methods with Mixed Input Embeddings and Loss Functions. We calculate the performance gap between each ablation design with the `Discretized` method.

7 Discussions

In this paper, we comprehensively evaluate how input representations and pretraining loss impact the zero-shot performance of generative EHR foundation models. Existing models have customized choices for the dataset, tokenizer, backbone, and training budget. We design a controlled study that aligns all these components to exclusively study their impact. We address limitations of current zero-shot evaluations that span a limited set of tasks. We propose `ZEHRO`, a comprehensive benchmark that includes classification, time-to-event, and regression tasks, with both one-step prediction and long-horizon simulation.

Our empirical results show that the `Continuous-fused` method favors regression tasks, while the `Discretized` method favors classification tasks. `MTPP` achieves the best performance in predicting large time intervals, which benefits its long-horizon time-to-event prediction. Our ablation studies show that no single combination of input embedding and pretraining loss is optimal across all tasks. Pretraining loss generally has a larger impact compared to input embeddings on zero-shot performance, and their correlation is task-dependent.

Limitations and future work. This study has several limitations. First, we evaluate only one representative method from each of three generative EHR foundation model families: `Discretized`, `Continuous-fused`, and `MTPP`. Future work could evaluate a broader range of methods and investigate how their input embeddings and loss functions affect downstream performance. Second, we use a relatively small Transformer backbone with approximately 40 million parameters, and our findings may not generalize to larger models. Future work could investigate how models with different input representations and pretraining objectives scale with model size. Finally, our ablation studies reveal task-dependent trade-offs among input representations and loss functions. Developing generative EHR foundation models that perform consistently well across diverse tasks remains an open challenge.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing, especially to check the grammar and word choice. We also used it to provide feedback for our figures. We have not used generative AI tools for literature review, generating datasets, or producing citation files. Additionally, we used generative AI tools for optimizing our code efficiency, like implementing the KV cache and continuous paging algorithm to accelerate the inference speed. We have reviewed all AI-assisted text and checked all related code for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Felix Draxler, Yang Meng, Kai Nelson, Lukas Laskowski, Yibo Yang, Theofanis Karaletsos, and Stephan Mandt. Transformers for mixed-type event sequences. *Advances in Neural Information Processing Systems*, 38:114948–114980, 2026.

Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, Geraud Krawezik, Francois Lanusse, Michael McCabe, Ruben Ohana, Liam Parker, et al. xval: a continuous numerical tokenization for scientific language models. *arXiv preprint arXiv:2310.02989*, 2023.

Lin Lawrence Guo, Santiago Eduardo Arciniegas, Joseph Jihyung Lee, Adam Paul Yan, George Tomlinson, Jason Fries, and Lillian Sung. Tokenization tradeoffs in structured ehr foundation models. *arXiv preprint arXiv:2603.15644*, 2026.

Zilin Jing, Vincent Jeanselme, Yuta Kobayashi, Simon A Lee, Chao Pang, Aparajita Kashyap, Yanwei Li, Xinzhuo Jiang, and Shalmali Joshi. One loss to rule them all: Marked time-to-event for structured ehr foundation models. *arXiv preprint arXiv:2602.00541*, 2026.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.

Inhyeok Lee, Luke Solo, Michael C Burkhart, Bashar Ramadan, William F Parker, and Brett K Beaulieu-Jones. Representation before training: A fixed-budget benchmark for generative medical event models. *arXiv preprint arXiv:2604.16775*, 2026.

Andrew J Loza, Jun Yup Kim, Shangzheng Song, Yihang Liu, Joseph JY Sung, R Andrew Taylor, and Dennis L Shung. multivariategpt: a decoder-only transformer for multivariate categorical and numeric data. *arXiv preprint arXiv:2505.21680*, 2025.

Kent A McCann, Ikgyu Shin, Huan Li, Davis White, Edward R Melnick, Mark S Iscoe, and Andrew J Loza. Continuous value tokenization improves medical event foundation models. *medRxiv*, pp. 2026–08, 2026.

Maria Elkjær Montgomery, Christian Igel, Mikkel Odgaard, Martin Sillesen, and Mads Nielsen. How should transformers encode numeric values in electronic health records? *arXiv preprint arXiv:2607.01391*, 2026.

Chao Pang, Xinzhuo Jiang, Krishna S. Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer (eds.), *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 239–260. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/pang21a.html.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Pawel Renc, Michal K Grzeszczyk, Nassim Oufattole, Deirdre Goode, Yugang Jia, Szymon Bieganski, Matthew BA McDermott, Jaroslaw Was, Anthony E Samir, Jonathan W Cunningham, et al. Foundation model of electronic medical records for adaptive risk estimation. *GigaScience*, 14: giaf107, 2025.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, 2025.

Ethan Steinberg, Yizhe Xu, Jason Alan Fries, and Nigam Shah. MOTOR: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NialiwI2V6.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, et al. Generative medical event models improve with scale. *arXiv preprint arXiv:2508.12104*, 2025.

Sheng Zhang, Qin Liu, Naoto Usuyama, Cliff Wong, Tristan Naumann, and Hoifung Poon. Exploring scaling laws for ehr foundation models, 2025. URL https://arxiv.org/abs/2505.22964.

A APPENDIX

A.1 DATASET

Table 2: Patient demographics in the Institution dataset. Values are percentages (counts).

Characteristic	Institution % (n)
Sex	
Female	55.7 (3,744,185)
Male	44.0 (2,955,940)
Missing	0.3 (16,814)
Ethnicity	
Hispanic or Latino	8.3 (557,173)
Not Hispanic or Latino	16.6 (1,118,232)
Missing	75.1 (5,041,534)
Race	
Asian	1.2 (83,697)
Black or African American	5.9 (396,677)
Native Hawaiian or Other Pacific Islander	0.1 (6,534)
Native American or Alaskan Native	0.1 (7,443)
White	16.9 (1,133,364)
Other	14.1 (944,342)
Missing	61.7 (4,144,882)

Table 3: Annual event counts by data type in the Institution dataset. Values are means (standard deviations).

Data type	Institutional
Healthcare utilization (distinct events per year)	118.4 (162.1)
Conditions (occurrences per year)	98.9 (392.9)
Medications (occurrences per year)	84.3 (508.6)
Procedures (occurrences per year)	45.8 (222.7)
Lab tests (occurrences per year)	817.0 (4,823.0)

A.2 TRAINING

We use 4 NVIDIA B200 GPUs (180GB) to pretrain models on MIMIC-IV and 4 NVIDIA H100 GPUs (96GB) to pretrain models on Institution dataset. For Institution, we only pretrain on 20% patients to to approximately match the total number of clinical events in MIMIC-IV.

Table 4: Pre-training hyperparameters shared by all loss families and ablation arms. Every arm uses the same trunk, stream, packed windows and budget; only the tokenization, the loss and its output heads differ.

Group	Hyperparameter	Value
Architecture	Trunk	GPT-2 (pre-LN), no bias terms
	Layers / heads / width	6 / 12 / 768
	Context length	2,048 positions
	Positional encoding	RoPE
	Activation	GELU
	Dropout	0.1
	Backbone parameters	$\sim$ 40M
Budget	Steps	30,000
	Tokens per step	$64 \times 4 \times 2{,}048 = 2^{19}$ (batch $\times$ accumulation $\times$ context)
	Total tokens	≈ 15.7B
	Checkpoint selection	best validation loss (evaluated every 1,000 steps)
Optimization	Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.95$
	Peak / final learning rate	1.5×10^{-3} / 1.5×10^{-4}
	Schedule	linear warm-up for 5% of steps (1,500), cosine decay to the final rate
	Weight decay	0.05
	Gradient clipping	1.0

A.3 TEST-TIME SCALING

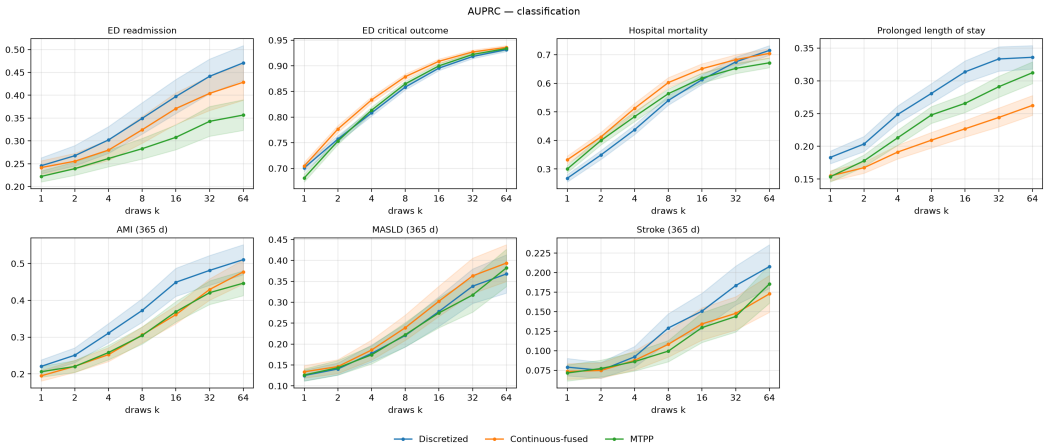


Figure 6: Test-time Scaling of Inference for Classification Tasks

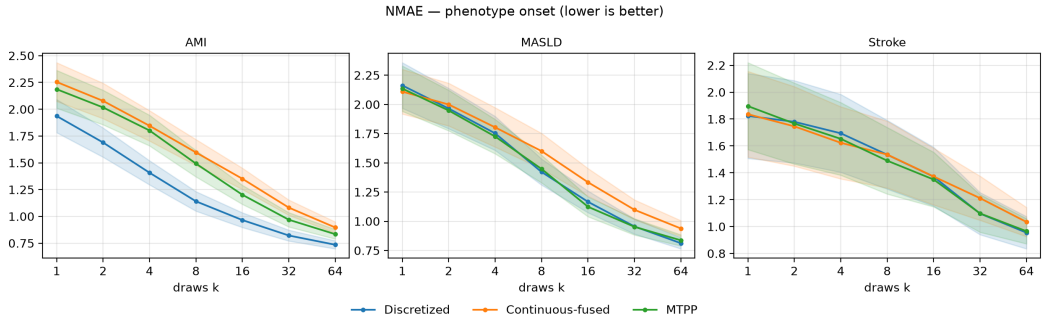


Figure 7: Test-time Scaling of Inference for Time-To-Event Tasks

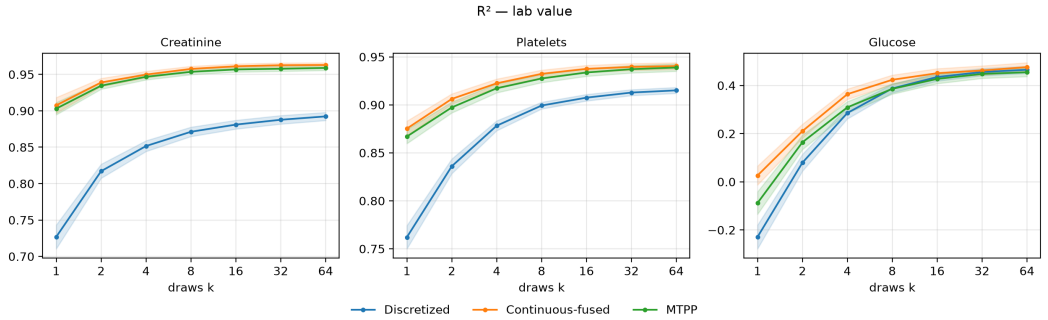


Figure 8: Test-time Scaling of Inference for Lab Regression Tasks

A.4 Ablation Study of Discretized Bins

Table 5: Effect of the number of discretization bins on zero-shot performance. Using 50 bins leads the regression value but loses on most classification tasks and time-to-even tasks. We choose 20 bins as it achieves a balanced performance across different tasks.

Task	10 bins	20 bins	50 bins
(a) Long-horizon classification, AUPRC ↑			
ED representation (72 h)	0.421 [0.361, 0.489]	**0.471** [0.397, 0.543]	0.432 [0.361, 0.506]
ED critical outcome (12 h)	0.930 [0.923, 0.937]	0.932 [0.924, 0.939]	**0.934** [0.928, 0.941]
Hospital mortality	0.706 [0.675, 0.735]	0.715 [0.684, 0.746]	**0.716** [0.687, 0.744]
Prolonged length of stay ($>$ 10 d)	0.245 [0.222, 0.270]	**0.336** [0.304, 0.374]	0.316 [0.286, 0.354]
Phenotype AMI (365 d)	0.432 [0.367, 0.503]	**0.511** [0.434, 0.589]	0.503 [0.431, 0.583]
Phenotype MASLD (365 d)	0.376 [0.281, 0.476]	0.368 [0.282, 0.461]	**0.397** [0.299, 0.497]
Phenotype stroke (365 d)	0.165 [0.122, 0.227]	**0.208** [0.159, 0.267]	0.189 [0.147, 0.240]
(b) Time prediction, NMAE ↓			
A2 next interval, true gap 1 min–1 h	**7.326** [7.023, 7.664]	14.898 [14.254, 15.625]	13.050 [12.496, 13.675]
A2 next interval, true gap 1 h–1 d	**5.436** [5.320, 5.561]	10.059 [9.808, 10.334]	9.749 [9.508, 9.993]
A2 next interval, true gap 1 d–30 d	**0.778** [0.769, 0.787]	1.586 [1.557, 1.614]	1.454 [1.431, 1.478]
A2 next interval, true gap 30 d–2 y	0.845 [0.843, 0.846]	**0.613** [0.611, 0.615]	0.630 [0.627, 0.632]
Phenotype AMI onset (positives)	**0.705** [0.650, 0.777]	0.735 [0.674, 0.816]	0.757 [0.686, 0.849]
Phenotype MASLD onset (positives)	0.827 [0.751, 0.935]	**0.812** [0.726, 0.915]	0.837 [0.752, 0.958]
Phenotype stroke onset (positives)	**0.913** [0.750, 1.185]	0.954 [0.779, 1.247]	0.924 [0.770, 1.184]
(c) Value prediction, R^2 ↑			
A3 next lab value (per-code z units, pooled)	0.528 [0.520, 0.535]	0.605 [0.597, 0.613]	**0.679** [0.670, 0.688]
Creatinine (24 h horizon)	0.744 [0.723, 0.768]	0.892 [0.882, 0.904]	**0.945** [0.938, 0.952]
Platelets (24 h horizon)	0.874 [0.864, 0.884]	0.915 [0.909, 0.921]	**0.934** [0.926, 0.940]
Glucose (24 h horizon)	0.439 [0.412, 0.462]	**0.465** [0.434, 0.492]	0.463 [0.427, 0.492]

A.5 Detailed Result

A.5.1 Detailed Performance for MIMIC-IV Dataset

Table 6: Absolute performance of the three loss families. Mean ± patient-clustered std; bold = best per row.

	Discrete	Continuous-fused	MTPP
One-step prediction			
A1 next diagnosis (AUPRC, top-100 codes)	$\mathbf{0.449 \pm 0.003}$	0.442 ± 0.003	0.432 ± 0.003
A2 next interval, true gap 1 min–1 h (NMAE) ↓	$\mathbf{14.898 \pm 0.353}$	24.332 ± 0.600	17.677 ± 0.443
A2 next interval, true gap 1 h–1 d (NMAE) ↓	$\mathbf{10.059 \pm 0.138}$	13.213 ± 0.183	11.224 ± 0.156
A2 next interval, true gap 1 d–30 d (NMAE) ↓	$\mathbf{1.586 \pm 0.015}$	2.393 ± 0.021	1.821 ± 0.018
A2 next interval, true gap 30 d–2 y (NMAE) ↓	0.613 ± 0.001	$\mathbf{0.581 \pm 0.001}$	0.605 ± 0.001
A2 next interval, geometric mean of the 4 gap bins (NMAE) ↓	$\mathbf{3.474 \pm 0.032}$	4.599 ± 0.043	3.846 ± 0.036
A3 next lab value (R^2, pooled z)	0.605 ± 0.004	$\mathbf{0.771 \pm 0.004}$	0.767 ± 0.005
Long-horizon simulation			
ED representation, 72 h (AUPRC)	$\mathbf{0.471 \pm 0.037}$	0.429 ± 0.039	0.357 ± 0.033
ED critical outcome, 12 h (AUPRC)	0.932 ± 0.004	$\mathbf{0.936 \pm 0.004}$	0.934 ± 0.004
Hospital mortality (AUPRC)	$\mathbf{0.715 \pm 0.016}$	0.705 ± 0.016	0.672 ± 0.016
Prolonged length of stay, >10 d (AUPRC)	$\mathbf{0.336 \pm 0.018}$	0.263 ± 0.016	0.312 ± 0.019
Phenotype AMI, 365 d (AUPRC)	$\mathbf{0.511 \pm 0.041}$	0.477 ± 0.034	0.447 ± 0.034
Phenotype MASLD, 365 d (AUPRC)	0.368 ± 0.047	$\mathbf{0.394 \pm 0.049}$	0.382 ± 0.046
Phenotype stroke, 365 d (AUPRC)	$\mathbf{0.208 \pm 0.027}$	0.173 ± 0.024	0.186 ± 0.025
Phenotype AMI onset, positives only (NMAE) ↓	$\mathbf{0.735 \pm 0.037}$	0.896 ± 0.050	0.834 ± 0.043
Phenotype MASLD onset, positives only (NMAE) ↓	$\mathbf{0.812 \pm 0.050}$	0.938 ± 0.068	0.837 ± 0.049
Phenotype stroke onset, positives only (NMAE) ↓	$\mathbf{0.954 \pm 0.121}$	1.034 ± 0.115	0.965 ± 0.096
Creatinine, 24 h (R^2)	0.892 ± 0.006	$\mathbf{0.963 \pm 0.003}$	0.959 ± 0.003
Platelets, 24 h (R^2)	0.915 ± 0.003	$\mathbf{0.940 \pm 0.004}$	0.939 ± 0.004
Glucose, 24 h (R^2)	0.465 ± 0.015	$\mathbf{0.476 \pm 0.018}$	0.455 ± 0.018

A.5.2 Detailed Performance for Ablation Studies

Table 7: Absolute performance of the representation × loss ablation. Discrete-gaussian = discrete input, continuous (Gaussian) loss; Continuous-categorical = continuous input, discretization loss. Mean ± patient-clustered std; bold = best per row.

	Discrete	Discrete-gaussian	Continuous-categorical	Continuous-fused
One-step prediction				
A1 next diagnosis (AUPRC, top-100 codes)	$\mathbf{0.449 \pm 0.003}$	0.438 ± 0.003	0.448 ± 0.003	0.442 ± 0.003
A2 next interval, true gap 1 min–1 h (NMAE) ↓	14.898 ± 0.353	24.357 ± 0.596	$\mathbf{14.237 \pm 0.337}$	24.332 ± 0.600
A2 next interval, true gap 1 h–1 d (NMAE) ↓	10.059 ± 0.138	13.679 ± 0.187	$\mathbf{9.903 \pm 0.134}$	13.213 ± 0.183
A2 next interval, true gap 1 d–30 d (NMAE) ↓	1.586 ± 0.015	2.548 ± 0.022	$\mathbf{1.538 \pm 0.014}$	2.393 ± 0.021
A2 next interval, true gap 30 d–2 y (NMAE) ↓	0.613 ± 0.001	$\mathbf{0.580 \pm 0.001}$	0.614 ± 0.001	0.581 ± 0.001
A2 next interval, geometric mean of the 4 gap bins (NMAE) ↓	3.474 ± 0.032	4.711 ± 0.043	$\mathbf{3.397 \pm 0.031}$	4.599 ± 0.043
A3 next lab value (R^2, pooled z)	0.605 ± 0.004	0.729 ± 0.004	0.620 ± 0.004	$\mathbf{0.771 \pm 0.004}$
Long-horizon simulation				
ED representation, 72 h (AUPRC)	$\mathbf{0.471 \pm 0.037}$	0.427 ± 0.039	0.461 ± 0.038	0.429 ± 0.039
ED critical outcome, 12 h (AUPRC)	0.932 ± 0.004	0.924 ± 0.004	0.928 ± 0.004	$\mathbf{0.936 \pm 0.004}$
Hospital mortality (AUPRC)	$\mathbf{0.715 \pm 0.016}$	0.679 ± 0.017	0.693 ± 0.017	0.705 ± 0.016
Prolonged length of stay, >10 d (AUPRC)	$\mathbf{0.336 \pm 0.018}$	0.250 ± 0.015	0.325 ± 0.019	0.263 ± 0.016
Phenotype AMI, 365 d (AUPRC)	$\mathbf{0.511 \pm 0.041}$	0.498 ± 0.032	0.490 ± 0.033	0.477 ± 0.034
Phenotype MASLD, 365 d (AUPRC)	0.368 ± 0.047	0.402 ± 0.049	$\mathbf{0.419 \pm 0.051}$	0.394 ± 0.049
Phenotype stroke, 365 d (AUPRC)	0.208 ± 0.027	0.194 ± 0.024	$\mathbf{0.217 \pm 0.026}$	0.173 ± 0.024
Phenotype AMI onset, positives only (NMAE) ↓	$\mathbf{0.735 \pm 0.037}$	0.923 ± 0.058	0.752 ± 0.036	0.896 ± 0.050
Phenotype MASLD onset, positives only (NMAE) ↓	$\mathbf{0.812 \pm 0.050}$	0.915 ± 0.057	0.840 ± 0.059	0.938 ± 0.068
Phenotype stroke onset, positives only (NMAE) ↓	$\mathbf{0.954 \pm 0.121}$	1.188 ± 0.155	1.011 ± 0.106	1.034 ± 0.115
Creatinine, 24 h (R^2)	0.892 ± 0.006	0.944 ± 0.003	0.895 ± 0.005	$\mathbf{0.963 \pm 0.003}$
Platelets, 24 h (R^2)	0.915 ± 0.003	0.928 ± 0.003	0.926 ± 0.003	$\mathbf{0.940 \pm 0.004}$
Glucose, 24 h (R^2)	0.465 ± 0.015	0.462 ± 0.018	0.460 ± 0.016	$\mathbf{0.476 \pm 0.018}$

Table 8: Table 3. Paired difference vs. Discrete for the ablation arms. Δ = arm $-$ Discrete in the metric's native units (NMAE: negative = better), $\pm$ paired std over 1000 draws; bold = $|\Delta| > 1.96$ std.

	+ continuous input embedding (Continuous-categorical)	+ continuous loss function (Discrete-gaussian)	+ continuous input embedding & loss function (Continuous-fused)
One-step prediction			
A1 next diagnosis (AUPRC, top-100 codes)	-0.000 ± 0.001	$\mathbf{-0.010 \pm 0.001}$	$\mathbf{-0.007 \pm 0.001}$
A2 next interval, true gap 1 min–1 h (NMAE) ↓	$\mathbf{-0.661 \pm 0.076}$	$\mathbf{+9.458 \pm 0.287}$	$\mathbf{+9.434 \pm 0.292}$
A2 next interval, true gap 1 h–1 d (NMAE) ↓	$\mathbf{-0.157 \pm 0.032}$	$\mathbf{+3.619 \pm 0.066}$	$\mathbf{+3.154 \pm 0.064}$
A2 next interval, true gap 1 d–30 d (NMAE) ↓	$\mathbf{-0.048 \pm 0.002}$	$\mathbf{+0.962 \pm 0.009}$	$\mathbf{+0.808 \pm 0.008}$
A2 next interval, true gap 30 d–2 y (NMAE) ↓	$\mathbf{+0.002 \pm 0.000}$	$\mathbf{-0.032 \pm 0.001}$	$\mathbf{-0.031 \pm 0.001}$
A2 next interval, geometric mean of the 4 gap bins (NMAE) ↓	$\mathbf{-0.076 \pm 0.007}$	$\mathbf{+1.237 \pm 0.016}$	$\mathbf{+1.125 \pm 0.016}$
A3 next lab value (R^2, pooled z)	$\mathbf{+0.015 \pm 0.001}$	$\mathbf{+0.124 \pm 0.003}$	$\mathbf{+0.167 \pm 0.003}$
Long-horizon simulation			
ED representation, 72 h (AUPRC)	-0.010 ± 0.009	$\mathbf{-0.044 \pm 0.012}$	$\mathbf{-0.042 \pm 0.011}$
ED critical outcome, 12 h (AUPRC)	-0.004 ± 0.002	$\mathbf{-0.007 \pm 0.002}$	$\mathbf{+0.004 \pm 0.002}$
Hospital mortality (AUPRC)	$\mathbf{-0.022 \pm 0.010}$	$\mathbf{-0.037 \pm 0.011}$	-0.010 ± 0.012
Prolonged length of stay, >10 d (AUPRC)	-0.011 ± 0.013	$\mathbf{-0.086 \pm 0.018}$	$\mathbf{-0.073 \pm 0.016}$
Phenotype AMI, 365 d (AUPRC)	-0.021 ± 0.017	-0.012 ± 0.018	-0.033 ± 0.018
Phenotype MASLD, 365 d (AUPRC)	$\mathbf{+0.052 \pm 0.021}$	$+0.035 \pm 0.018$	$+0.026 \pm 0.022$
Phenotype stroke, 365 d (AUPRC)	$+0.010 \pm 0.014$	-0.014 ± 0.017	-0.035 ± 0.018
Phenotype AMI onset, positives only (NMAE) ↓	$+0.017 \pm 0.022$	$\mathbf{+0.188 \pm 0.037}$	$\mathbf{+0.161 \pm 0.035}$
Phenotype MASLD onset, positives only (NMAE) ↓	$+0.028 \pm 0.042$	$+0.103 \pm 0.055$	$\mathbf{+0.126 \pm 0.054}$
Phenotype stroke onset, positives only (NMAE) ↓	$+0.057 \pm 0.060$	$\mathbf{+0.233 \pm 0.070}$	$+0.080 \pm 0.061$
Creatinine, 24 h (R^2)	$+0.002 \pm 0.002$	$\mathbf{+0.052 \pm 0.006}$	$\mathbf{+0.071 \pm 0.007}$
Platelets, 24 h (R^2)	$\mathbf{+0.011 \pm 0.001}$	$\mathbf{+0.013 \pm 0.002}$	$\mathbf{+0.025 \pm 0.003}$
Glucose, 24 h (R^2)	-0.005 ± 0.007	-0.003 ± 0.008	$+0.011 \pm 0.009$

A.5.3 Detailed Results in Institution Dataset

Table 9: Absolute performance of the three loss families on the Institution held-out split (zero-shot generic evaluation). Mean $\pm$ std over 1000 row-level bootstrap draws; bold = best per row.

	Discrete	Continuous-fused	MTPP
One-step prediction			
A1 next diagnosis (AUPRC, top-100 codes)	$\mathbf{0.513 \pm 0.014}$	0.510 ± 0.014	0.463 ± 0.016
A2 next interval, true gap 1 min–1 h (NMAE) ↓	10.149 ± 0.403	$\mathbf{10.095 \pm 0.410}$	108.622 ± 2.488
A2 next interval, true gap 1 h–1 d (NMAE) ↓	$\mathbf{1.909 \pm 0.037}$	1.986 ± 0.043	10.757 ± 0.186
A2 next interval, true gap 1 d–30 d (NMAE) ↓	$\mathbf{0.994 \pm 0.008}$	1.045 ± 0.010	2.357 ± 0.025
A2 next interval, true gap 30 d–2 y (NMAE) ↓	0.780 ± 0.002	0.786 ± 0.002	$\mathbf{0.688 \pm 0.003}$
A2 next interval, geometric mean of all gap bins (NMAE) ↓	$\mathbf{1.708 \pm 0.016}$	1.737 ± 0.017	4.432 ± 0.028
A3 next lab value (R^2, pooled z)	0.484 ± 0.021	$\mathbf{0.587 \pm 0.017}$	0.564 ± 0.017
Long-horizon simulation			
Hospital mortality (AUPRC)	$\mathbf{0.831 \pm 0.011}$	0.809 ± 0.013	0.757 ± 0.013
Prolonged length of stay, >5 d (AUPRC)	$\mathbf{0.840 \pm 0.010}$	0.807 ± 0.011	0.773 ± 0.012
30-day readmission (AUPRC)	$\mathbf{0.725 \pm 0.013}$	0.707 ± 0.013	0.670 ± 0.014
Phenotype AMI, 365 d (AUPRC)	$\mathbf{0.254 \pm 0.018}$	0.225 ± 0.017	0.251 ± 0.019
Phenotype MASLD, 365 d (AUPRC)	0.515 ± 0.014	$\mathbf{0.549 \pm 0.015}$	0.508 ± 0.014
Phenotype stroke, 365 d (AUPRC)	$\mathbf{0.248 \pm 0.019}$	0.246 ± 0.019	0.248 ± 0.018
Phenotype AMI onset (NMAE) ↓	$\mathbf{0.953 \pm 0.056}$	1.089 ± 0.065	0.967 ± 0.055
Phenotype MASLD onset (NMAE) ↓	0.886 ± 0.030	$\mathbf{0.831 \pm 0.028}$	0.837 ± 0.029
Phenotype stroke onset (NMAE) ↓	0.798 ± 0.040	0.869 ± 0.044	$\mathbf{0.785 \pm 0.040}$
Creatinine, 24 h (R^2)	0.794 ± 0.023	$\mathbf{0.889 \pm 0.024}$	0.864 ± 0.024
Platelets, 24 h (R^2)	0.671 ± 0.042	$\mathbf{0.746 \pm 0.041}$	0.650 ± 0.054
Glucose, 24 h (R^2)	0.259 ± 0.031	$\mathbf{0.453 \pm 0.025}$	0.313 ± 0.030

A.6 Empirical Distribution of Time Interval and Lab Value

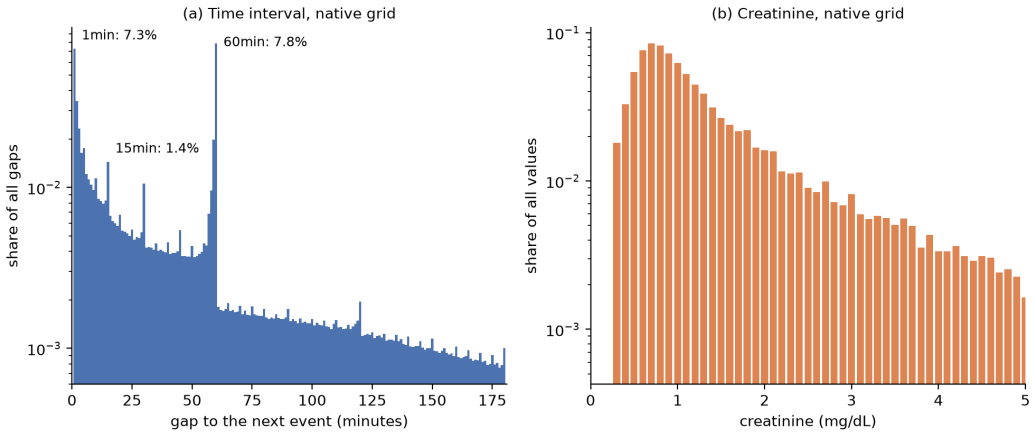


Figure 9: Empirical Distribution of Time Intervals and Lab Value (Creatinine) in MIMIC-IV

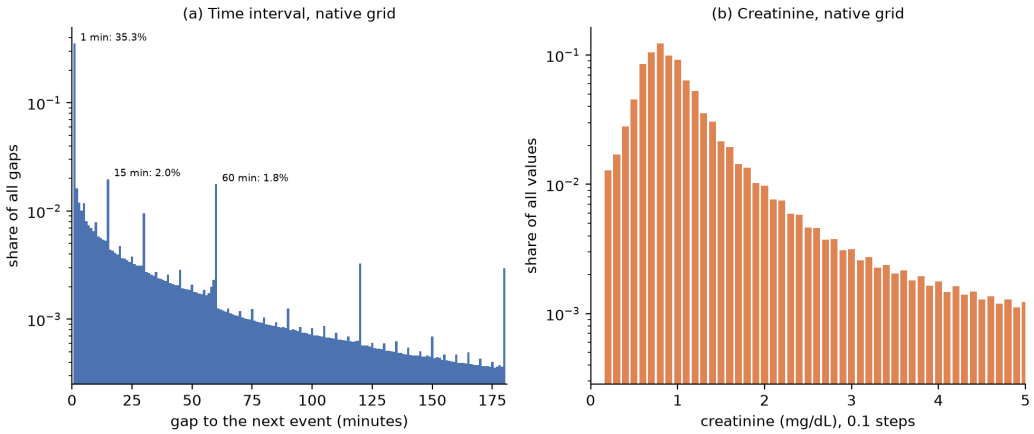


Figure 10: Empirical Distribution of Time Intervals and Lab Value (Creatinine) in Institution Dataset