FORMATTING INSTRUCTIONS FOR ICLR 2027 CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models for electronic health records aim to synthesize plausible future clinical trajectories conditioned on a patient's prior history. These approaches support a range of downstream applications, including risk stratification and phenotype discovery. However, most existing methods generate trajectories at a predetermined resolution - on the order of hours or days - making long-horizon forecasting computational expensive and error-prone. The few approaches that incorporate multiple temporal resolutions scale poorly to multiple outcomes of interest and require costly training procedures. Thus, we present Multiple Resolution Generation of Electronic Health Records (MuRGEHR – pronounced merger), a framework for the efficient generation at multiple resolution over long horizons without sacrificing fidelity. MuRGEHR models the EHR time series as a hierarchical sequence of tokens representing different temporal resolutions. The model first constructs a coarse scaffold, which then serves as conditioning for synthesizing progressively finer resolutions. Across three datasets and multiple prediction horizons, this hierarchical strategy yields consistent gains in computational efficiency while remaining competitive with regards to fidelity on downstream clinical tasks.

1 INTRODUCTION

Clinical risk models that predict a patient's future state from their past history can improve care by enabling early intervention, better disease management, and more effective treatment options (Tomašev et al., 2019; Muralitharan et al., 2021). Such models are typically constructed for a given pre-specified endpoint, often leveraging a pretrained model that is fine-tuned for each downstream task (Li et al., 2020; Rasmy et al., 2021; Pang et al., 2021).

Generative models of electronic health records (EHR) offer an alternative way to sample a patient's prospective clinical trajectory without having to specify predefined endpoints (Renc et al., 2024; Steinberg et al., 2024; Kraljevic et al., 2024). During inference, the model auto-regressively produces multiple future trajectories for a given patient via Monte Carlo rollout over a specified horizon. The likelihood of any observation within that horizon is then estimated from its empirical frequency across the sampled trajectories.

Event-stream generative models represent EHR as a temporally interleaved sequence of event tokens (Renc et al., 2024; 2025), making the cost of generating a patient trajectory scale with the density of events within the horizon. By contrast, models that aggregate at the visit level or impose a fixed temporal granularity decouple cost from event density; their computational burden depends primarily on the length of the horizon, but the temporal resolution is predetermined and cannot adapt to the underlying data (Redekop et al., 2025). This makes it difficult to capture the varying temporal dynamics of events. Moreover, a fine-resolution model is expensive to rollout for long-horizon tasks, and a coarse (low-resolution) model cannot predict tasks that occur at a smaller resolution. One approach is to pretrain a separate model per temporal resolution and fuse the outputs from these models for a given downstream task (Oufattole et al.). However, pretraining separate generative models for each resolution is expensive and inference is still costly because one needs to rollout the trajectories at each resolution and then combine the results afterwards.

To address these limitations, we developed a Multiple Resolution Generation of Electronic Health Records (MuRGEHR – pronounced merger) method - a single generative model over nested resolu-

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.

tions. The model conditions on a multi-resolution past context and generates multi-resolution future trajectories in a coarse-to-fine manner such that the finer resolutions are conditioned on the coarser ones. We utilize attention masking so that each coarse bin and its descendants are self-contained, which allows for generating the minimum number of tokens for any given horizon as well as parallel refinement once coarse trajectories are generated. We demonstrate performance benefits on downstream clinical tasks and quantify the computational benefits.

Our contributions are:

- We introduce a hierarchical framework for modeling health trajectories that captures both coarse (low-resolution) and fine (high-resolution) temporal structure.
- We generate each future trajectory at multiple temporal resolutions, refining from low-to-high resolution, enabling a parsimonious representation in which clinically salient intervals can be synthesized at high resolution while less informative periods are modeled more coarsely.
- We demonstrate that the MuRGEHR is a promising alternative to event stream models, achieving competitive performance on clinical tasks while using significantly less compute resources by using a greedy approach that minimizes tokens and latency during generation for a given time horizon.
- We perform ablations to probe the effectiveness of our design choices, and we evaluate the uncertainty of our generations by leveraging our hierarchical generation process.

2 Related Works

Prior work in EHR modeling typically use sequence based models, such as transformers, and differ mainly in how time is discretized. Visit-level models aggregate events that occur in a single visit and model patient records as a sequence of encounters and some notion of time (Li et al., 2020; Rasmy et al., 2021; Pang et al., 2021). However, these implementations do not model intra-visit time, which may span weeks, and are unable to predict when future events occur. Event-level models treat every event as a token, forming an event stream with interleaved time difference tokens (Renc et al., 2025; 2024; Yang et al., 2023). While these models can do autoregressive rollout to generate future trajectories, the sequence length of a patient trajectory reflects event density as opposed a clinical window of interest. Moreover, in order to determine what happens at a specific time in the future, these models must generate all events that predate it, which is costly and allows for error to compound over long sequences.

Grid models fix a time axis in advance and predict occupancy of codes per bin. This makes the budget a function of the horizon and allows for the modeling of multiple resolutions. MoRGEN utilizes a mixture of experts over multiple temporal resolutions to capture the varying temporal dynamics between codes and demonstrates performant results in clinical tasks (Oufattole et al.). However, it requires many pre-trained generative models over different temporal resolutions and rolling out several clinical trajectories for each model. Additionally, these models have yet to demonstrate the ability to achieve competitive performance compared with event stream models without supervision.

A parallel line of work focuses on time-to-event models, which estimate each code's hazard over a discrete time interval grid and integrate the resulting hazard function in closed form. MOTOR utilizes a piecewise-exponential head on top of a pretrained encoder that outputs a per-time-interval hazard for each code, (Steinberg et al., 2024). This approach is computationally efficient because it avoids ordered rollouts, but its expressiveness is constrained by the piecewise-constant hazard grid, which limits fine-grained temporal predictions.

While hierarchical generation is well established outside of EHR generation, they are often used to achieve fine-grained outcomes more efficiently and are not evaluated on their coarse level readouts. Funnel-Transformer (Dai et al., 2020) and Hourglass Transformers (Nawrot et al., 2022) downsample tokens internally to save compute. MegaByte (Yu et al., 2023) first aggregates the sequence into patches and learns within and across patches. VAR (Tian et al., 2024) recasts image generation as residual predictions between resolutions. N-HiTS (Challu et al., 2023) varies the sampling rate to create the hierarchy, and Pyraformer (Liu et al., 2021) creates the hierarchy using convolutions. In all of these methods, performance is not evaluated at the level of the coarse hierarchy. For Hourglass, MegaByte and Pyraformer the coarse temporal resolution corresponds to pooled hidden states with

no observable attached; for VAR they are residual quantizations of a latent feature map, recoverable only by summing across all scales before decoding; for N-HiTS the levels are additive frequency components, so no individual stack output corresponds to a coarse forecast of the target. In sum, none of these models evaluate on their coarse readouts and utilize the coarse readouts during inference for performance benefits. This represents a missed opportunity as coarse readouts are cheap to obtain and contain useful information on longer time scales.

3 METHODS

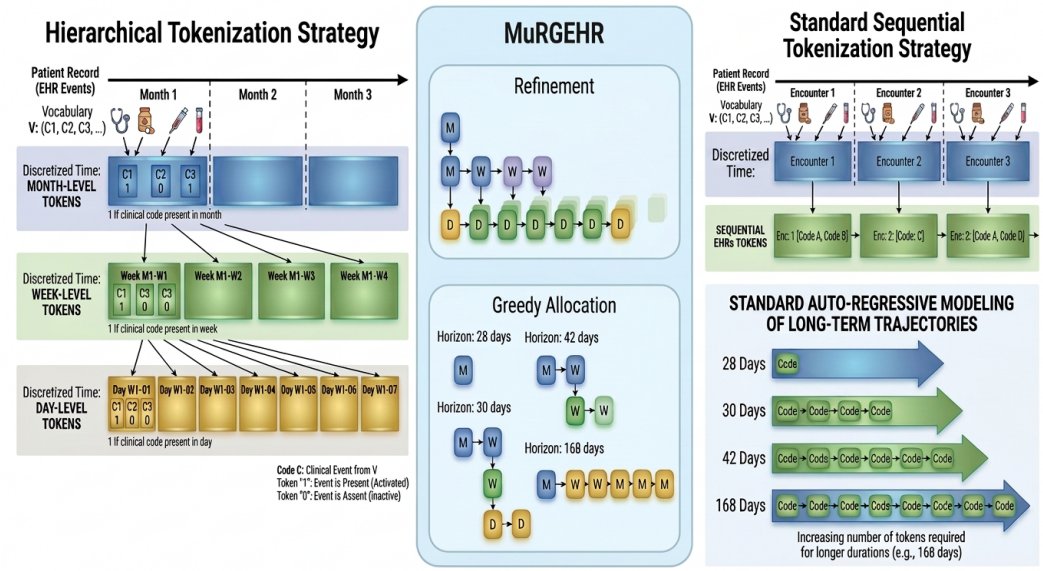


Figure 1: MuRGEHR's fixed sequence length per horizon with refinement and greedy decoding options compared to standard autoregressive models, where sequence length scales with event density.

3.1 TOKENIZATION

A patient record is a set of events over a fixed vocabulary of V clinical codes, where each event is encoded as (code, timestamp). We discretize time into n bins of different temporal lengths and represent the record as a binary tensor $x \in \{0,1\}^{V\times n}$, where $x_{c,b} = 1$ iff code c occurs in bin b as shown in Figure 1. This allows us to keep a fixed sequence length independent of event density of a patient.

We choose to tokenize each time bin formed as the pooled embedding of the codes that appear in that bin, following:

$$h_b \;=\; \underbrace{\mathrm{Sum}\big(\{\, e_c : x_{c,b} = 1 \,\}\big) + \mathbb{1}[\,b \text{ empty}\,]\, e_{\varnothing}}_{\text{pooled content}} \;+ p_b \;+\; r_{\rho(b)}, \tag{1}$$

where e_c is a learned code embedding, $e_{\varnothing}$ is a learned embedding for a bin with no active code (whose sum-pool would otherwise be exactly $\mathbf{0}$), p_b is a temporal positional embedding, and r_ρ a resolution embedding. All embedding are $\in \mathbb{R}^{256}$. Position embeddings are a learned absolute embedding for the ordinal index of b within its own resolution. Resolution embeddings are a learned embedding that encapsulates the resolution of bin b (month, week or year).

3.2 HIERARCHICAL SETUP

We begin by representing a patient timeline anchored at discharge using three temporal resolutions: month (28 days), week (7 days), and day corresponding to Coarse (C), Intermediate (I), and Fine (F). We use a 504 day window with 280 days of past context pre-discharge and 224 days of future post-discharge at all three resolutions flattened into a sequence.

Let ρ represent a given resolution $\in \{C, I, F\}$ for coarse, intermediate, and fine and index it $\{0, 1, 2\}$. Let n_ρ be the number of bins within each resolution over the 504 day context window length; i.e., $n_C = 18$, $n_I = 72$, $n_F = 504$ bins over the window. Each coarse bin together with its descendants (defined to be the bins at a finer resolution that span the same time) form a *time chunk*. For example, a month level token has 4 associated week bins and a week covers 7 day bins.

This nested hierarchy allows us to formulate predictions as a residual prediction task and allows the model to answer a given horizon query using a greedy allocation method that minimizes the number of tokens generated per trajectory. Once a coarse (month) token is generated, the week head produces a correction that is added to the parent's (month's) upsampled logits, rather than producing an absolute logit itself. A similar approach is applied to generating a day-level token, given a corresponding month and week-level token.

During training, coarse tokens attend to past history at all resolutions and any previous generated coarse token. Intermediate tokens attend to its parent coarse token and any previous intermediate token sharing the same coarse parent, and fine can attend to its parent coarse and intermediate token (one per resolution) and any previous fine token sharing the same coarse parent.

Residual Predictions Since the different resolutions of the grid are nested, a finer head does not need to predict the entire logit that may be present within its coarser parent. Thus, we formulate this as a residual prediction task where the coarsest resolution emits absolute logits and each finer one emits a correction. Thus, a fine logit is a coarse logit plus an intermediate correction plus a fine correction. The parents logit is evenly distributed among its children and then we then learn a residual correction to this value. This residual prediction mirrors the way we design the attention mask, perform readout, and do inference.

Architecture. MuRGEHR is a decoder-only transformers with 4 layers, 4 attention heads, hidden dimension $d = 256$ and feed-forward width $4d$. The output head (one per resolution) is a two layer mlp to V independent sigmoids per bin, with one head per resolution. We optimize using cross-entropy per resolution weighted by the number of bins.

Loss Function Let $\ell^\rho_{c,b}$ the logit the model emits for code c in bin b. Only observed cells of the generated region are scored: $\mathcal{O}_\rho = \{(c, b) : b \geq \beta_\rho,\ o^\rho_b = 1\}$, where o^ρ_b indicates that bin b lies within the patient's follow-up. Then

$$\mathcal{L}(\theta) \;=\; \sum_\rho w_\rho \, \frac{1}{|\mathcal{O}_\rho|} \sum_{(c,b)\in\mathcal{O}_\rho} \mathrm{BCE}\Big(\sigma\big(\ell^\rho_{c,b}\big),\, x^\rho_{c,b}\Big), \qquad w_\rho = \frac{R\, n_\rho}{\sum_{\rho'} n_{\rho'}}. \tag{2}$$

The per-resolution weights, w_ρ, are proportional to bin count rather than uniform because this would skew the model towards the coarser resolutions due to the fewer tokens. Because the month grid has $28\times$ fewer bins than the day grid, each month bin would receive $28\times$ the gradient of each day bin if using uniform weighting. Setting $w_\rho \propto n_\rho$ cancels that factor and gives every bin equal weight, normalized to sum to the number of resolutions R: $w = (0.09, 0.36, 2.55)$ for $(18, 72, 504)$ bins.

3.3 INFERENCE

Given a horizon of interest, we use a greedy algorithm to minimize the number of tokens generated until the horizon is reached (Fig. 1). This means greedily allocating months, then weeks, and finally days until the horizon of interest is reached and not refining the resolutions for coarse allocations. To mirror the training setup, the coarse parent of any finer resolution will also be generated to act as conditioning but will not be used for evaluated. Thus, for 28 days, 30 days, 6 week, and 6 month evals, MuRGEHR will evaluate using 1 month token, 1 month and 2 day tokens, 1 month token and 2 week tokens, and 6 month tokens respectively, but for 30 day and 6 week, generate an additional month and week token and month token respectively.

4 EXPERIMENTAL SETUP

4.1 DATASETS

We evaluate our approach on two publicly available datasets—MIMIC-IV (Johnson et al., 2024) and EHR-Shot (Wornow et al., 2023)—as well as a private health-system dataset (Hospital A) comprising heart-failure patients receiving care at large tertiary centers. Dataset statistics are located in Appendix A.1 (Table A.1, A.2). Performance is assessed across a set of clinically relevant endpoints (elevated creatinine; decreased levels of hematocrit, hemoglobin, leukocyte (WBC), platelet; presence of death, readmission, and ICU and ED admission when applicable) commonly used for risk stratification and management in heart disease.

4.2 BASELINES

We compare MuRGEHR to three autoregressive baseline approaches:

1. We train single resolution models using only month tokens, week tokens, and day tokens. Each model only sees one resolution of past patient history and is asked to generate trajectories at the same temporal resolution. The architecture is the same as MuRGEHR, with the exception of the multi-resolution heads. We refer to these as month, week, and day models corresponding to their resolution.
2. We also train an event stream model where each token can be a code or time interval; i.e., observed time intervals are part of the model's vocabulary with the smallest time delta being a minute. The architecture is the same as MuRGEHR, with the exception of the multi-resolution heads. We refer to this model as the event stream model.
3. ETHOS is retrained on each dataset with a matched sequence and sequence length (Renc et al., 2024).

Additional implementation details are located in Appendix C.

4.3 EVALUATION

We generate 100 trajectories for each query. We report the AUROC. The AUPRC and Expected Calibration Error (ECE) are in the Supplementary Data sections A.3 (Table A.7, A.8, A.9) and A.4 (Table A.10, A.11, **??**). We also report the leave-one-out trajectory standard jacknife error.

5 RESULTS

We define the single outcome task to be whether or not an abnormal lab or a given event occurred within a time horizon: $\text{fires}(A, H) = [\tau_A < H]$, at $H \in \{30, 168\}$ days.

At a fixed sample budget, MuRGEHR demonstrates strong performance compared to the best performing baseline (Tables 1, 2, 3). While ETHOS has a higher macro-AUC and macro-AUPRC, MuRGEHR remains competitive at a small fraction of the generation cost; the average per discharge latency and tokens of MuRGEHR for a single trajectory with no speedups (KV caching (Pope et al., 2023) or vLLm (Kwon et al., 2023)) is significantly lower than all models except the coarsest resolution one (Tables 1, 2, 3).

At the same compute budget (i.e., given the same compute time) MuRGEHR matches or exceeds ETHOS for AUC and AUPRC (see Appendix A.3 Table A.7, A.8, A.9) with the exception of AUPRC on EHRSHOT, which is the smallest dataset. We compare them at equal wall-clock time, where MuRGEHR is allowed to generate as many trajectories as it can given exactly the amount of time it took Ethos to generate its 100 trajectories for the held out cohort. Even though ETHOS ran with vLLm (both used fp16), MuRGEHR (no caching) is able to generate 6x-10x more trajectories given the same amount of time. Moreover, when we compare performance, MuRGEHR begins to outperform Ethos on AUROC and AUPRC. Additional horizons are in Appendix A.2.1 (Table A.3, A.4, A.5) and the same conclusions hold.

# of gen. traj	month 100	week 100	day 100	event stream 100	*ETHOS* 100	MuRGEHR 100	MuRGEHR 643
30 days							
Creat.	–	–	**.944** (.002)	.930 (.003)	.929 (.003)	.934 (.003)	.943 (.001)
Hct	–	–	**.891** (.002)	.877 (.002)	.887 (.002)	.885 (.002)	**.891** (.001)
Hgb	–	–	.885 (.002)	.877 (.003)	.882 (.002)	.882 (.002)	**.893** (.001)
WBC	–	–	.865 (.001)	.855 (.001)	.859 (.002)	.859 (.001)	**.866** (.001)
Plt	–	–	.946 (.004)	.931 (.003)	.942 (.003)	.941 (.003)	**.953** (.001)
Death	–	–	.850 (.006)	.848 (.006)	.868 (.006)	.848 (.006)	**.872** (.002)
Readm.	–	–	.694 (.002)	.699 (.002)	**.703** (.002)	.688 (.002)	.700 (.001)
ICU	–	–	.724 (.006)	.705 (.007)	.729 (.005)	.716 (.007)	**.732** (.003)
ED	–	–	.713 (.002)	.718 (.003)	**.722** (.002)	.704 (.002)	.713 (.001)
macro	–	–	.835 (.001)	.827 (.001)	.836 (.001)	.829 (.001)	**.840** (.001)
gen. tokens / traj	–	–	28.7	39.6	120.1	4.9	4.9
ms / traj	–	–	169.7	167.4	414.5	26.8	26.8
6 months							
Creat.	.921 (.002)	.921 (.002)	.930 (.001)	.927 (.002)	.920 (.002)	.930 (.001)	**.933** (<.001)
Hct	.846 (.001)	.850 (.001)	.862 (.001)	.860 (.001)	.863 (.001)	.864 (.001)	**.868** (<.001)
Hgb	.850 (.001)	.853 (.001)	.864 (.001)	.865 (.001)	.865 (.001)	.867 (.001)	**.871** (<.001)
WBC	.841 (.001)	.840 (.001)	.851 (.001)	.851 (.001)	**.859** (.001)	.839 (.001)	.843 (<.001)
Plt	.896 (.002)	.897 (.002)	.908 (.002)	.898 (.003)	.905 (.003)	.907 (.002)	**.913** (.001)
Death	.828 (.002)	.833 (.002)	.822 (.002)	.842 (.003)	.841 (.002)	.841 (.002)	**.848** (.001)
Readm.	.691 (.002)	.703 (.001)	.706 (.001)	.715 (.002)	**.727** (.002)	.707 (.001)	.711 (.001)
ICU	.712 (.003)	.732 (.003)	.732 (.002)	.725 (.003)	.731 (.004)	.728 (.003)	**.736** (.001)
ED	.720 (.002)	.720 (.001)	.728 (.001)	.733 (.002)	**.746** (.002)	.727 (.001)	.730 (.001)
macro	.812 (.001)	.816 (.001)	.823 (.001)	.824 (.001)	**.828** (.001)	.823 (.001)	**.828** (<.001)
gen. tokens / traj	5.5	17.6	129.9	39.6	120.1	5.0	5.0
ms / traj	23.0	113.0	744.4	167.4	414.5	36.4	36.4

Table 1: Per-outcome AUROC and generation cost at 30 days (top) and 6 months (bottom), MIMIC.

# of gen. traj	month 100	week 100	day 100	event stream 100	*ETHOS* 100	MuRGEHR 100	MuRGEHR 707
30 days							
Creat.	–	–	.807 (.006)	.902 (.009)	**.915** (.007)	.889 (.011)	.888 (.004)
Hct	–	–	.836 (.003)	.824 (.003)	.837 (.005)	.840 (.004)	**.853** (.002)
Hgb	–	–	.834 (.004)	.826 (.004)	.832 (.006)	.845 (.006)	**.854** (.002)
WBC	–	–	.735 (.004)	.951 (.003)	**.963** (.001)	.915 (.004)	.918 (.002)
Plt	–	–	.865 (.002)	.908 (.003)	**.948** (.002)	.917 (.003)	.923 (.001)
Death	–	–	.721 (.048)	.639 (.026)	.757 (.033)	.741 (.030)	**.835** (.012)
Readm.	–	–	.690 (.003)	.651 (.003)	.656 (.004)	.696 (.005)	**.703** (.002)
ED	–	–	.590 (.010)	.609 (.006)	**.649** (.006)	.615 (.007)	.627 (.004)
macro	–	–	.760 (.006)	.789 (.004)	.820 (.004)	.807 (.004)	**.825** (.002)
gen. tokens / traj	–	–	29.4	95.8	167.3	4.7	4.7
ms / traj	–	–	115.4	379.8	623.9	20.2	20.2
6 months							
Creat.	.888 (.008)	.875 (.004)	.811 (.003)	.913 (.004)	.912 (.006)	.903 (.005)	**.918** (.002)
Hct	.754 (.006)	.818 (.003)	.803 (.002)	.800 (.003)	.822 (.003)	.829 (.003)	**.834** (.001)
Hgb	.761 (.006)	.814 (.003)	.809 (.003)	.802 (.003)	.812 (.003)	.827 (.003)	**.834** (.001)
WBC	.832 (.006)	.692 (.003)	.627 (.004)	.914 (.004)	**.935** (.003)	.836 (.005)	.846 (.001)
Plt	.888 (.003)	.895 (.001)	.844 (.002)	.884 (.002)	**.920** (.002)	.892 (.002)	.897 (.001)
Death	.767 (.016)	.761 (.013)	.721 (.017)	.717 (.013)	.776 (.018)	.813 (.010)	**.816** (.004)
Readm.	.611 (.003)	.671 (.002)	.664 (.003)	.683 (.002)	.668 (.004)	**.697** (.003)	**.697** (.001)
ED	.630 (.003)	.651 (.004)	.624 (.006)	.636 (.004)	**.680** (.004)	.659 (.004)	.668 (.002)
macro	.766 (.002)	.772 (.002)	.738 (.002)	.794 (.002)	**.815** (.002)	.807 (.002)	.814 (.001)
gen. tokens / traj	5.8	23.3	154.7	95.8	167.3	5.2	5.2
ms / traj	22.2	95.1	592.4	379.8	623.9	21.5	21.5

Table 2: Per-outcome AUROC and generation cost at 30 days (top) and 6 months (bottom), EHR-SHOT.

# of gen. traj	month 100	week 100	day 100	event stream 100	*ETHOS* 100	**MuRGEHR** 100	**MuRGEHR** 1098
30 days							
Creat.	–	–	.953 (.002)	.952 (.002)	**.963** (.001)	.959 (.001)	**.963** (.000)
Hct	–	–	.907 (.001)	.906 (.001)	.910 (.001)	.909 (.001)	**.913** (.000)
Hgb	–	–	.920 (.001)	.920 (.001)	.923 (.001)	.926 (.001)	**.928** (.000)
WBC	–	–	.874 (.001)	.869 (.001)	.875 (.001)	.880 (.001)	**.883** (.000)
Plt	–	–	.919 (.002)	.918 (.002)	.921 (.002)	.924 (.002)	**.931** (.001)
Death	–	–	.797 (.005)	.834 (.005)	.835 (.004)	.828 (.005)	**.840** (.001)
Readm.	–	–	.757 (.001)	.745 (.001)	.751 (.003)	.765 (.001)	**.770** (.000)
macro	–	–	.875 (.001)	.878 (.001)	.883 (.001)	.884 (.001)	**.890** (.000)
gen. tokens / traj.	–	–	29.4	46.2	187.9	4.9	4.9
ms / traj.	–	–	53.8	77.0	570.1	15.9	15.9
6 months							
Creat.	.943 (.001)	.942 (.001)	.937 (.001)	.938 (.001)	.946 (.001)	.944 (.001)	**.948** (.000)
Hct	.845 (.001)	.851 (.001)	.854 (.001)	.856 (.001)	.859 (.001)	.859 (.001)	**.863** (.000)
Hgb	.865 (.001)	.870 (.001)	.871 (.001)	.876 (.001)	.878 (.001)	.879 (.001)	**.882** (.000)
WBC	.848 (.001)	.850 (.001)	.853 (.001)	.851 (.001)	.849 (.001)	.853 (.001)	**.856** (.000)
Plt	.875 (.002)	.872 (.001)	.874 (.001)	.875 (.002)	.877 (.002)	.879 (.001)	**.883** (.000)
Death	.771 (.002)	.770 (.002)	.763 (.002)	**.792** (.003)	.782 (.003)	.781 (.002)	.787 (.001)
Readm.	.710 (.001)	.724 (.001)	.725 (.001)	.725 (.001)	.729 (.001)	.731 (.001)	**.738** (.000)
macro	.837 (.000)	.840 (.000)	.840 (.000)	.845 (.001)	.846 (.001)	.847 (.000)	**.851** (.000)
gen. tokens / traj.	5.2	21.4	147.9	46.2	187.9	5.4	5.4
ms / traj.	11.1	42.7	295.9	77.0	570.1	18.8	18.8

Table 3: Per-outcome AUROC and generation cost per patient at 30 days (top) and 6 months (bottom), Hospital A.

	MuRGEHR	MuRGEHR w/o residual	MuRGEHR w/o inter. and fine past	MuRGEHR w/o attn mask
30 days				
Creat.	.914 (.005)	.914 (.006)	.915 (.004)	.923 (.006)
Hct	.879 (.003)	.850 (.004)	.880 (.003)	.879 (.004)
Hgb	.868 (.003)	.850 (.004)	.871 (.004)	.875 (.004)
WBC	.849 (.002)	.845 (.002)	.847 (.002)	.850 (.002)
Plt	.943 (.005)	.915 (.004)	.933 (.004)	.936 (.004)
Death	.822 (.009)	.832 (.009)	.817 (.009)	.820 (.009)
Readm.	.659 (.003)	.637 (.002)	.657 (.003)	.667 (.003)
ICU	.678 (.010)	.641 (.009)	.675 (.010)	.690 (.009)
ED	.694 (.004)	.661 (.003)	.685 (.003)	.687 (.003)
macro	.812 (.002)	.794 (.002)	.809 (.002)	.814 (.002)
tokens	4.9	4.9	4.8	4.9
ms / traj	21.6 (±3.6)	22.0 (±4.0)	13.6 (±1.8)	32.2 (±5.8)
6 months				
Creat.	.919 (.002)	.905 (.003)	.913 (.002)	.911 (.002)
Hct	.860 (.002)	.848 (.002)	.861 (.002)	.870 (.002)
Hgb	.864 (.002)	.851 (.002)	.862 (.001)	.866 (.002)
WBC	.839 (.001)	.839 (.001)	.836 (.001)	.842 (.001)
Plt	.914 (.003)	.900 (.003)	.908 (.002)	.909 (.003)
Death	.820 (.003)	.825 (.003)	.815 (.003)	.816 (.003)
Readm.	.692 (.002)	.679 (.002)	.693 (.002)	.696 (.002)
ICU	.725 (.004)	.719 (.004)	.733 (.004)	.731 (.004)
ED	.718 (.002)	.700 (.002)	.716 (.002)	.721 (.002)
macro	.817 (.001)	.807 (.001)	.815 (.001)	.818 (.001)
tokens	5.1	5.1	5.0	5.1
ms / traj	22.9 (±7.6)	20.9 (±6.1)	13.3 (±4.2)	29.1 (±8.2)

Table 4: Ablation Results

5.1 Ablations

We perform three ablations on a 50k subset of MIMIC and evaluate on a 5k matched prevalence subset of the held out portion. We first remove the residual prediction head and instead force each resolution to output its own absolute logits. We also choose to evaluate whether we need multi-

	28 days				6 months			
	90% threshold		75% threshold		90% threshold		75% threshold	
	confident ↑	unconfident ↓	confident ↑	unconfident ↓	confident ↑	unconfident ↓	confident ↑	unconfident ↓
align. month↔week	.812	.798	.789	.797	.817	.668	.826	.734
align. all three res.	.824	.761	.823	.794	.816	.655	.827	.709

Table 5: Uncertainty signals, macro AUROC on the kept and on the deferred discharges at 28 days and 6 months.

resolution past patient history, or whether we can just use the coarse resolution, which would save memory and compute. Lastly, we ablate the per-chunk attention mask compared with full attention over every token. Results are shown in Table 4. Additional horizons are shown in Appendix A.2.2 (Table A.6).

The residual predictions from finer resolutions greatly improves the predictions with the exception of death. This suggests that the residual aids in performance of commonly occurring events and may not help with these rare event outcomes.

The second ablation restricts the coarse resolution to attend only to the coarse past, rather than the past at every resolution. While there is a slight drop off in performance at longer horizons, the compute budget is decreased significantly from 331 tokens of past conditioning to just 11 tokens. We also achieve latency performance benefits. This suggest that it may not be necessary to condition on all resolutions.

Lastly, we evaluate the use of the sparse, per-chunk attention mask with a full causal attention mask. This greatly increases latency during inference. The results suggest that the performance gains are marginal, with much larger compute costs.

5.2 Uncertainty Analysis

We stratify discharges by identifying cases where the model exhibits uncertainty. To do this, we exploit the multi-resolution structure of our predictions. Each coarse-resolution token can be expanded into its constituent fine-resolution tokens, and we expect these resolutions to be mutually consistent. For example, if a code is active on day 3, it should also appear in week 1 and month 1. Because this consistency is not explicitly enforced during training, any disagreement between resolutions serves as a proxy for model uncertainty.

For each trajectory and resolution pair, we compute the fraction of (code, parent-cell) pairs in which the parent and its OR-pooled children disagree, normalized by the number of pairs where either is active. We measure this disagreement between week and month resolutions—avoiding day-level generation for computational efficiency and also compute an aggregate disagreement score across all resolutions. Using these uncertainty estimates, we report AUROC separately for the most confident and least confident discharges in Table 5, evaluated on the greedy-allocated query set.

Using this ranking, MuRGEHR achieves improved performance on its confident predictions compared to its least confident predictions with the exception of 75% threshold at 28 days using align month and week. This is likely due to the small sampling from month and week pairs (4 per trajectory) as we see the performance shift drastically when evaluating 6 months at the same threshold.

6 Conclusion

We introduce MuRGEHR, a generative model over nested multi-resolution EHR data. By tokenizing each time bin, we fix the sequence length for every given patient, allowing the three resolutions to share a sequence as opposed to event stream models where event density may cause significantly longer sequences over the same time period. The unified framework allows information to propagate across time points, and the masking scheme allows for efficient answering of queries by emitting the coarsest token that answers it. Across all datasets, MuRGEHR performs competitively against single resolution grid models, an event stream baseline, and ETHOS on single-outcome tasks over various time horizons. We show the compute efficacy of MuRGEHR compared to the other autoregressive methods in terms of generated tokens and time, and given equal wall-time, demonstrate stronger

performance. Lastly, we perform ablations to motivate the various design choices and show that we can identify uncertain predictions for further analysis or increased trajectory generation.

AI USE STATEMENT

In this work, we used generative AI tools to design or provide feedback on research methodology or experiments and implement methods. We have not used generative AI tools for generating synthetic datasets, help in the development of theoretical models or conceptual framework, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, and interpret results are not applicable to this work.

Additionally, we have used generative AI tools for drafting and editing of the manuscript, figure generation. We have reviewed all AI assisted work. LLM generated code was reviewed for correctness. Generated figures were reviewed for correctness. We reviewed all AI assisted writing and have edited it in accordance with our work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the details to preprocess the data and the corresponding code to run MuRGEHR in this anonymized repo: `https://anonymous.4open.science/r/murgehr-9BBD/README.md`.

REFERENCES

Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 6989–6997, 2023.

Zihang Dai, Guokun Lai, Yiming Yang, and Quoc Le. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. *Advances in neural information processing systems*, 33:4271–4282, 2020.

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`. Version 3.1.

Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*, 2021.

Sankavi Muralitharan, Walter Nelson, Shuang Di, Michael McGillion, PJ Devereaux, Neil Grant Barr, and Jeremy Petch. Machine learning–based early warning systems for clinical deterioration: systematic scoping review. *Journal of medical Internet research*, 23(2):e25187, 2021.

Piotr Nawrot, Szymon Tworkowski, Michał Tyrolski, Łukasz Kaiser, Yuhuai Wu, Christian Szegedy, and Henryk Michalewski. Hierarchical transformers are more efficient language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 1559–1571, 2022.

Nassim Oufattole, Matthew BA McDermott, and Collin Stultz. Morgen: Mixture-of-resolutions generative forecasting for irregularly sampled medical time-series data. In *Forty-third International Conference on Machine Learning*.

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 *Machine learning for health*, pp. 239–260. PMLR, 2021.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of machine learning and systems*, 5:606–624, 2023.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

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*, 32(12):1833–1842, 2025.

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, 2024.

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.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. In *International Conference on Learning Representations*, volume 2024, pp. 41038–41077, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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, 2023.

Lili Yu, Dániel Simig, Colin Flaherty, Armen Aghajanyan, Luke Zettlemoyer, and Mike Lewis. Megabyte: Predicting million-byte sequences with multiscale transformers. *Advances in Neural Information Processing Systems*, 36:78808–78823, 2023.

A Appendix

A.1 Dataset Statistics

	MIMIC-IV				EHR-SHOT				Hospital A			
	train	val	held-out	**all**	train	val	held-out	**all**	train	val	held-out	**all**
Patients	146,531	3,832	3,906	**154,269**	1,957	1,876	1,884	**5,717**	96,072	2,513	2,533	**101,118**
Events (M)	26.99	0.68	0.73	**28.40**	0.83	0.69	0.68	**2.20**	30.99	0.81	0.82	**32.63**
Discharges	377,505	9,592	10,640	**397,737**	2,513	2,091	2,079	**6,683**	562,362	14,637	15,045	**592,044**
Vocabulary V	282				430				74			

Table A.1: Dataset Statistics

	MIMIC-IV				EHR-SHOT				Hospital A			
Outcome	28d	30d	42d	168d	28d	30d	42d	168d	28d	30d	42d	168d
Creatinine	.0428	.0439	.0520	.0929	.0322	.0327	.0375	.0688	.1018	.1043	.1113	.1433
Hematocrit	.0618	.0638	.0754	.1305	.0837	.0871	.1044	.1770	.1237	.1252	.1351	.1902
Hemoglobin	.0603	.0628	.0751	.1321	.0688	.0712	.0875	.1573	.1297	.1318	.1424	.1994
Leukocyte	.1461	.1516	.1775	.2867	.0698	.0712	.0818	.1236	.1966	.2016	.2225	.3226
Platelet	.0399	.0417	.0483	.0743	.1236	.1251	.1361	.1837	.0816	.0834	.0903	.1328
Death	.0315	.0331	.0430	.1037	.0178	.0183	.0231	.0693	.0362	.0379	.0464	.0966
Readmission	.2795	.2894	.3294	.5071	.1688	.1746	.2116	.3458	.3682	.3773	.4226	.6009
ICU admission	.0439	.0460	.0567	.1179	–	–	–	–	–	–	–	–
ED visit	.2074	.2161	.2466	.4047	.1255	.1299	.1578	.2727	–	–	–	–

Table A.2: Per-code held-out prevalences across datasets.

A.2 Additional Horizons

A.2.1 Benchmarking

	month	week	day	event stream	*ETHOS*	MuRGEHR	MuRGEHR
# of gen. traj	100	100	100	100	100	100	643
28 days							
Creat.	.925 (.004)	.928 (.003)	**.944** (.003)	.930 (.003)	.928 (.003)	.931 (.003)	.942 (.001)
Hct	.873 (.003)	.884 (.002)	.891 (.002)	.878 (.003)	.888 (.003)	.888 (.002)	**.892** (.001)
Hgb	.875 (.003)	.877 (.002)	.884 (.002)	.877 (.003)	.883 (.002)	.882 (.002)	**.892** (.001)
WBC	.852 (.001)	.859 (.001)	**.865** (.001)	.856 (.001)	.858 (.002)	.857 (.002)	**.865** (.001)
Plt	.938 (.004)	.943 (.003)	.948 (.004)	.934 (.003)	.946 (.003)	.944 (.003)	**.954** (.002)
Death	.837 (.006)	.863 (.006)	.850 (.006)	.848 (.006)	.870 (.006)	.855 (.007)	**.877** (.002)
Readm.	.675 (.003)	.690 (.002)	.695 (.002)	.699 (.002)	**.703** (.002)	.693 (.002)	.700 (.001)
ICU	.688 (.008)	.711 (.007)	.725 (.006)	.704 (.007)	.728 (.006)	.707 (.007)	**.731** (.003)
ED	.698 (.002)	.708 (.002)	.713 (.003)	.716 (.003)	**.720** (.002)	.703 (.003)	.712 (.001)
macro	.818 (.001)	.829 (.001)	.835 (.001)	.827 (.001)	.836 (.001)	.829 (.001)	**.840** (.001)
gen. tokens / traj	2.0	4.8	26.9	39.6	120.1	1.0	1.0
ms / traj	4.8	16.6	119.3	167.4	414.5	5.1	5.1
42 days							
Creat.	–	.928 (.003)	**.942** (.002)	.930 (.002)	.930 (.003)	.935 (.003)	.941 (.001)
Hct	–	.876 (.001)	**.884** (.001)	.870 (.002)	.878 (.002)	.877 (.002)	**.884** (.001)
Hgb	–	.877 (.002)	.883 (.001)	.874 (.002)	.879 (.002)	.883 (.002)	**.890** (.001)
WBC	–	.856 (.001)	**.862** (.001)	.857 (.001)	.858 (.002)	.857 (.001)	**.862** (<.001)
Plt	–	.938 (.003)	.940 (.003)	.930 (.003)	.941 (.003)	.935 (.003)	**.947** (.001)
Death	–	.860 (.004)	.852 (.004)	.852 (.006)	.868 (.005)	.854 (.005)	**.869** (.002)
Readm.	–	.697 (.002)	.697 (.002)	.707 (.002)	**.712** (.002)	.693 (.002)	.702 (.001)
ICU	–	.718 (.005)	.725 (.005)	.711 (.006)	**.735** (.005)	.693 (.006)	.731 (.003)
ED	–	.715 (.002)	.717 (.002)	.725 (.002)	**.731** (.002)	.706 (.002)	.719 (.001)
macro	–	.829 (.001)	.834 (.001)	.828 (.001)	.837 (.001)	.826 (.001)	**.838** (<.001)
gen. tokens / traj	–	6.5	39.3	39.6	120.1	3.9	3.9
ms / traj	–	36.8	187.9	167.4	414.5	19.9	19.9

Table A.3: Per-outcome AUROC and generation cost per patient at 28 days (top) and 42 days (bottom), MIMIC.

	month	week	day	event stream	*ETHOS*	MuRGEHR	MuRGEHR
# of gen. traj	100	100	100	100	100	100	707
28 days							
Creat.	.792 (.002)	.864 (.011)	.807 (.006)	.902 (.010)	**.918** (.008)	.899 (.011)	.895 (.004)
Hct	.707 (.010)	**.854** (.005)	.835 (.003)	.819 (.003)	.832 (.005)	.847 (.004)	.851 (.002)
Hgb	.716 (.012)	.850 (.005)	.834 (.004)	.825 (.004)	.834 (.006)	.842 (.005)	**.854** (.002)
WBC	.911 (.009)	.811 (.005)	.739 (.004)	.951 (.003)	**.962** (.001)	.919 (.005)	.916 (.002)
Plt	.901 (.005)	.922 (.002)	.863 (.002)	.911 (.003)	**.951** (.002)	.919 (.003)	.924 (.001)
Death	.684 (.023)	.639 (.037)	.699 (.046)	.642 (.031)	.770 (.034)	.743 (.026)	**.830** (.012)
Readm.	.612 (.005)	.683 (.004)	.695 (.003)	.651 (.003)	.655 (.004)	.702 (.005)	**.705** (.002)
ED	.592 (.007)	.643 (.011)	.593 (.011)	.603 (.006)	**.646** (.007)	.601 (.009)	.624 (.004)
macro	.739 (.004)	.783 (.005)	.758 (.006)	.788 (.004)	.821 (.005)	.809 (.004)	**.825** (.002)
gen. tokens / traj	1.0	4.0	27.4	95.8	167.3	1.0	1.0
ms / traj	5.5	15.0	106.8	379.8	623.9	4.7	4.7
42 days							
Creat.	–	.868 (.009)	.804 (.005)	.911 (.007)	**.923** (.006)	.896 (.009)	.902 (.003)
Hct	–	.836 (.004)	.821 (.003)	.814 (.004)	.835 (.004)	.832 (.004)	**.843** (.002)
Hgb	–	.839 (.004)	.826 (.003)	.818 (.004)	.828 (.005)	.846 (.004)	**.851** (.002)
WBC	–	.772 (.004)	.706 (.004)	.950 (.003)	**.963** (.003)	.901 (.005)	.909 (.002)
Plt	–	.919 (.002)	.861 (.002)	.907 (.003)	**.945** (.002)	.923 (.002)	.923 (.001)
Death	–	.686 (.034)	.723 (.042)	.689 (.025)	.752 (.029)	.794 (.025)	**.850** (.009)
Readm.	–	.681 (.003)	.682 (.003)	.672 (.003)	.668 (.004)	.695 (.004)	**.702** (.002)
ED	–	.633 (.008)	.599 (.008)	.613 (.005)	**.655** (.005)	.624 (.007)	.635 (.003)
macro	–	.779 (.005)	.753 (.006)	.797 (.004)	.821 (.004)	.814 (.003)	**.827** (.001)
gen. tokens / traj	–	6.0	41.0	95.8	167.3	3.8	3.8
ms / traj	–	22.6	200.3	379.8	623.9	22.8	22.8

Table A.4: Per-outcome AUROC and generation cost at 28 days and 42 days (top to bottom), EHR-SHOT.

	month	week	day	event stream	*ETHOS*	**MuRGEHR**	**MuRGEHR**
# of gen. traj	100	100	100	100	100	100	1098
28 days							
Creat.	.955 (.001)	.956 (.001)	.956 (.002)	.955 (.002)	.964 (.001)	.962 (.001)	**.965** (.000)
Hct	.887 (.001)	.900 (.001)	.908 (.001)	.907 (.001)	.911 (.001)	.909 (.001)	**.914** (.000)
Hgb	.904 (.001)	.913 (.001)	.921 (.001)	.921 (.001)	.924 (.001)	.925 (.001)	**.929** (.000)
WBC	.865 (.001)	.874 (.001)	.876 (.001)	.871 (.001)	.876 (.001)	.882 (.001)	**.884** (.000)
Plt	.912 (.002)	.918 (.002)	.922 (.002)	.921 (.002)	.924 (.002)	.928 (.002)	**.933** (.001)
Death	.806 (.005)	.810 (.004)	.795 (.005)	.835 (.005)	.836 (.004)	.820 (.004)	**.840** (.001)
Readm.	.728 (.001)	.751 (.001)	.759 (.001)	.747 (.001)	.752 (.003)	.766 (.001)	**.773** (.000)
macro	.865 (.001)	.875 (.001)	.877 (.001)	.879 (.001)	.884 (.001)	.885 (.001)	**.891** (.000)
gen. tokens / traj.	1	3.9	27.5	46.2	187.9	1	1
ms / traj.	2.4	7.9	50.1	77.0	570.1	3.2	3.2
42 days							
Creat.	–	.954 (.001)	.951 (.001)	.951 (.001)	**.962** (.001)	.958 (.001)	**.962** (.000)
Hct	–	.892 (.001)	.900 (.001)	.898 (.001)	.903 (.001)	.903 (.001)	**.907** (.000)
Hgb	–	.905 (.001)	.913 (.001)	.912 (.001)	.916 (.001)	.917 (.001)	**.921** (.000)
WBC	–	.870 (.001)	.871 (.001)	.865 (.001)	.869 (.001)	.874 (.001)	**.879** (.000)
Plt	–	.909 (.002)	.910 (.002)	.910 (.002)	.915 (.002)	.915 (.002)	**.923** (.001)
Death	–	.805 (.004)	.790 (.003)	.830 (.004)	.830 (.004)	.817 (.003)	**.834** (.001)
Readm.	–	.747 (.001)	.752 (.001)	.743 (.001)	.750 (.002)	.758 (.001)	**.764** (.000)
macro	–	.869 (.001)	.870 (.001)	.873 (.001)	.878 (.001)	.877 (.001)	**.884** (.000)
gen. tokens / traj.	–	5.7	40.9	46.2	187.9	3.9	3.9
ms / traj.	–	11.4	74.8	77.0	570.1	12.6	12.6

Table A.5: Per-outcome AUROC and generation cost per trajectory at 28 days (top) and 42 days (bottom), Hospital A.

A.2.2 Ablation Results

	MuRGEHR	MuRGEHR w/o residual	MuRGEHR w/o inter. and fine past	MuRGEHR w/o attn mask
28 days				
Creat.	.915 (.005)	.906 (.005)	.916 (.005)	.916 (.005)
Hct	.883 (.003)	.855 (.005)	.883 (.003)	.882 (.003)
Hgb	.871 (.004)	.847 (.005)	.872 (.004)	.872 (.003)
WBC	.850 (.002)	.844 (.002)	.846 (.002)	.850 (.002)
Plt	.939 (.005)	.927 (.005)	.944 (.006)	.944 (.004)
Death	.816 (.008)	.847 (.008)	.820 (.009)	.829 (.009)
Readm.	.659 (.003)	.640 (.002)	.659 (.003)	.668 (.003)
ICU	.660 (.011)	.636 (.010)	.659 (.011)	.694 (.011)
ED	.697 (.003)	.663 (.003)	.691 (.003)	.688 (.003)
macro	.810 (.002)	.796 (.002)	.810 (.002)	.816 (.002)
tokens	1.0	1.0	1.0	1.0
ms / traj	4.4 (±0.7)	6.2 (±0.7)	2.6 (±0.2)	5.6 (±0.3)
6 weeks				
Creat.	.912 (.006)	.905 (.006)	.918 (.006)	.918 (.005)
Hct	.871 (.003)	.846 (.004)	.869 (.003)	.870 (.004)
Hgb	.867 (.004)	.845 (.004)	.867 (.003)	.866 (.003)
WBC	.842 (.002)	.833 (.002)	.837 (.002)	.846 (.002)
Plt	.935 (.006)	.927 (.005)	.933 (.005)	.932 (.005)
Death	.803 (.007)	.811 (.007)	.798 (.007)	.824 (.008)
Readm.	.668 (.003)	.643 (.002)	.664 (.003)	.681 (.003)
ICU	.675 (.009)	.646 (.008)	.679 (.008)	.691 (.009)
ED	.695 (.003)	.666 (.003)	.690 (.003)	.700 (.004)
macro	.807 (.002)	.791 (.002)	.806 (.002)	.814 (.002)
tokens	3.9	3.9	3.9	3.9
ms / traj	16.3 (±2.0)	18.0 (±3.4)	11.5 (±2.1)	23.7 (±3.5)

Table A.6: Ablation Results at additional horizons (28 days and 6 weeks).

A.3 AUPRC

# of gen. traj	month 100	week 100	day 100	event stream 100	*ETHOS* 100	MuRGEHR 100	MuRGEHR 643
28 days							
Creat.	.466 (.008)	.519 (.007)	.532 (.007)	.427 (.004)	.466 (.008)	.510 (.007)	**.537** (.005)
Hct	.428 (.004)	.445 (.004)	.455 (.003)	.413 (.004)	.439 (.004)	.446 (.005)	**.464** (.002)
Hgb	.376 (.004)	.388 (.004)	.413 (.004)	.357 (.004)	.380 (.005)	.403 (.005)	**.417** (.003)
WBC	.572 (.002)	.588 (.002)	**.599** (.002)	.552 (.002)	.576 (.004)	.585 (.002)	.595 (.001)
Plt	.639 (.005)	.683 (.005)	.676 (.004)	.647 (.005)	**.685** (.005)	.664 (.004)	.677 (.002)
Death	.163 (.006)	.245 (.006)	.284 (.006)	.257 (.006)	**.307** (.008)	.211 (.007)	.227 (.004)
Readm.	.477 (.003)	.501 (.002)	.505 (.002)	.518 (.002)	**.532** (.002)	.497 (.003)	.510 (.001)
ICU	.094 (.004)	.112 (.004)	**.120** (.004)	.107 (.003)	.115 (.004)	.101 (.006)	.117 (.002)
ED	.422 (.003)	.455 (.003)	.460 (.003)	.465 (.002)	**.473** (.002)	.448 (.003)	.460 (.002)
macro	.404 (.002)	.437 (.002)	**.449** (.002)	.416 (.001)	.442 (.002)	.429 (.002)	.445 (.001)
gen. tokens / traj	2.0	4.8	26.9	39.6	120.1	1.0	1.0
ms / traj	4.8	16.6	119.3	167.4	414.5	5.1	5.1
30 days							
Creat.	–	–	.541 (.005)	.435 (.004)	.476 (.008)	.542 (.008)	**.551** (.005)
Hct	–	–	.459 (.003)	.417 (.003)	.444 (.004)	.445 (.004)	**.468** (.002)
Hgb	–	–	.422 (.004)	.368 (.004)	.387 (.005)	.401 (.005)	**.423** (.002)
WBC	–	–	**.606** (.002)	.558 (.002)	.584 (.004)	.598 (.002)	.605 (.002)
Plt	–	–	.671 (.004)	.644 (.005)	**.683** (.005)	.664 (.006)	.681 (.002)
Death	–	–	.283 (.006)	.261 (.006)	**.310** (.007)	.220 (.007)	.235 (.004)
Readm.	–	–	.513 (.002)	.528 (.002)	**.542** (.002)	.509 (.002)	.524 (.001)
ICU	–	–	**.123** (.004)	.112 (.003)	.120 (.003)	.110 (.003)	.122 (.003)
ED	–	–	.469 (.003)	.474 (.003)	**.485** (.002)	.455 (.003)	.470 (.004)
macro	–	–	**.454** (.001)	.422 (.001)	.448 (.002)	.438 (.002)	.453 (.001)
gen. tokens / traj	–	–	28.7	39.6	120.1	4.9	4.9
ms / traj	–	–	169.7	167.4	414.5	26.8	26.8
42 days							
Creat.	–	.572 (.006)	.593 (.005)	.493 (.004)	.507 (.009)	.581 (.007)	**.604** (.003)
Hct	–	.457 (.003)	.465 (.003)	.445 (.003)	.452 (.004)	.465 (.004)	**.480** (.002)
Hgb	–	.426 (.003)	.445 (.003)	.407 (.003)	.409 (.004)	.437 (.004)	**.451** (.002)
WBC	–	.624 (.002)	.636 (.002)	.603 (.002)	.616 (.003)	.626 (.002)	**.637** (.001)
Plt	–	.688 (.003)	.683 (.004)	.665 (.005)	**.693** (.004)	.674 (.003)	.688 (.002)
Death	–	.273 (.006)	.298 (.006)	.289 (.006)	**.328** (.006)	.247 (.006)	.267 (.004)
Readm.	–	.565 (.002)	.561 (.002)	.580 (.002)	**.594** (.002)	.561 (.002)	.574 (.001)
ICU	–	.147 (.004)	**.163** (.006)	.145 (.004)	.154 (.004)	.127 (.005)	.145 (.002)
ED	–	.512 (.002)	.512 (.002)	.520 (.002)	**.532** (.002)	.501 (.003)	.518 (.002)
macro	–	.474 (.001)	.484 (.002)	.461 (.001)	.476 (.002)	.469 (.001)	**.485** (.001)
gen. tokens / traj	–	6.5	39.3	39.6	120.1	3.9	3.9
ms / traj	–	36.8	187.9	167.4	414.5	19.9	19.9
6 months							
Creat.	.688 (.003)	.703 (.003)	.706 (.002)	.684 (.004)	.661 (.008)	.703 (.003)	**.713** (.001)
Hct	.490 (.003)	.512 (.003)	.513 (.003)	.524 (.003)	.524 (.003)	.535 (.003)	**.542** (.001)
Hgb	.486 (.003)	.501 (.003)	.504 (.003)	.511 (.003)	.501 (.003)	.521 (.003)	**.534** (.002)
WBC	.712 (.002)	.717 (.002)	.729 (.002)	.722 (.001)	**.733** (.002)	.721 (.002)	.730 (.001)
Plt	.638 (.003)	.653 (.003)	.644 (.004)	.663 (.003)	**.674** (.004)	.647 (.003)	.658 (.002)
Death	.362 (.003)	.399 (.004)	.371 (.003)	.432 (.004)	**.440** (.004)	.412 (.005)	.433 (.002)
Readm.	.716 (.002)	.722 (.002)	.724 (.001)	.737 (.001)	**.745** (.001)	.728 (.001)	.737 (.001)
ICU	.280 (.005)	**.313** (.003)	.312 (.004)	.298 (.003)	**.313** (.004)	.297 (.004)	.306 (.005)
ED	.664 (.001)	.671 (.002)	.676 (.002)	.681 (.002)	**.695** (.002)	.675 (.001)	.683 (.001)
macro	.560 (.001)	.577 (.001)	.575 (.001)	.584 (.001)	.587 (.002)	.582 (.001)	**.593** (.001)
gen. tokens / traj	5.5	17.6	129.9	39.6	120.1	5.0	5.0
ms / traj	23.0	113.0	744.4	167.4	414.5	36.4	36.4

Table A.7: Per-outcome AUPRC (↑) and generation cost per trajectory at 28 days, 30 days, 42 days and 6 months, MIMIC.

	month	week	day	event stream	*ETHOS*	MuRGEHR	MuRGEHR
# of gen. traj	100	100	100	100	100	100	691
28 days							
Creat.	.302 (.011)	.198 (.011)	.149 (.011)	.362 (.015)	**.444** (.014)	.332 (.037)	.304 (.018)
Hct	.206 (.006)	**.368** (.012)	.331 (.014)	.337 (.011)	.351 (.009)	.339 (.014)	.355 (.008)
Hgb	.173 (.005)	.303 (.011)	.261 (.009)	**.337** (.008)	.270 (.007)	.314 (.015)	.320 (.013)
WBC	.644 (.010)	.417 (.010)	.206 (.012)	.565 (.007)	**.705** (.019)	.647 (.012)	.659 (.006)
Plt	.636 (.006)	.704 (.006)	.486 (.010)	.686 (.010)	**.772** (.009)	.698 (.011)	.701 (.006)
Death	**.130** (.017)	.036 (.012)	.040 (.009)	.037 (.007)	.079 (.010)	.059 (.011)	.079 (.014)
Readm.	.248 (.004)	.282 (.004)	.291 (.005)	.240 (.003)	.266 (.005)	.321 (.010)	**.330** (.006)
ED	.155 (.002)	.191 (.011)	.163 (.008)	.199 (.007)	**.231** (.009)	.171 (.006)	.180 (.004)
macro	.312 (.003)	.313 (.004)	.241 (.004)	.345 (.003)	**.390** (.005)	.360 (.006)	.366 (.004)
gen. tokens / traj	1.0	4.0	27.4	95.8	167.3	1.0	1.0
ms / traj	5.5	15.0	106.8	379.8	623.9	4.7	4.7
30 days							
Creat.	–	–	.139 (.008)	.359 (.016)	**.440** (.015)	.293 (.016)	.301 (.019)
Hct	–	–	.339 (.013)	.343 (.009)	.357 (.009)	.347 (.014)	**.363** (.010)
Hgb	–	–	.266 (.008)	**.340** (.010)	.272 (.006)	.302 (.015)	.323 (.011)
WBC	–	–	.208 (.011)	.570 (.007)	**.710** (.020)	.643 (.009)	.664 (.006)
Plt	–	–	.494 (.011)	.680 (.009)	**.768** (.009)	.694 (.010)	.701 (.006)
Death	–	–	.041 (.009)	.038 (.005)	.076 (.011)	.051 (.019)	**.089** (.022)
Readm.	–	–	.296 (.004)	.248 (.003)	.273 (.005)	.336 (.008)	**.340** (.005)
ED	–	–	.167 (.009)	.208 (.006)	**.243** (.007)	.177 (.006)	.187 (.005)
macro	–	–	.244 (.004)	.348 (.003)	**.392** (.004)	.355 (.005)	.371 (.005)
gen. tokens / traj	–	–	29.4	95.8	167.3	4.7	4.7
ms / traj	–	–	115.4	379.8	623.9	20.2	20.2
42 days							
Creat.	–	.243 (.016)	.167 (.014)	.439 (.013)	**.481** (.013)	.349 (.020)	.352 (.012)
Hct	–	.384 (.008)	.357 (.009)	.368 (.008)	.390 (.008)	.374 (.018)	**.403** (.006)
Hgb	–	.335 (.012)	.306 (.010)	.357 (.008)	.311 (.007)	.364 (.014)	**.374** (.010)
WBC	–	.384 (.007)	.196 (.011)	.619 (.006)	**.721** (.018)	.641 (.010)	.663 (.005)
Plt	–	.714 (.007)	.513 (.010)	.686 (.008)	**.765** (.009)	.716 (.010)	.716 (.004)
Death	–	.060 (.016)	.053 (.010)	.052 (.006)	.086 (.011)	.076 (.014)	**.100** (.014)
Readm.	–	.358 (.005)	.350 (.005)	.315 (.003)	.324 (.004)	.389 (.008)	**.401** (.006)
ED	–	.230 (.009)	.205 (.008)	.246 (.005)	**.276** (.006)	.221 (.007)	.229 (.005)
macro	–	.338 (.004)	.268 (.004)	.385 (.002)	**.419** (.004)	.391 (.005)	.405 (.003)
gen. tokens / traj	–	6.0	41.0	95.8	167.3	3.8	3.8
ms / traj	–	22.6	200.3	379.8	623.9	22.8	22.8
6 months							
Creat.	.560 (.012)	.424 (.011)	.244 (.015)	.617 (.006)	**.636** (.007)	.500 (.013)	.516 (.014)
Hct	.405 (.005)	.487 (.006)	.456 (.008)	.458 (.004)	.497 (.005)	.511 (.009)	**.533** (.006)
Hgb	.383 (.004)	.444 (.008)	.423 (.009)	.437 (.005)	.433 (.006)	.485 (.009)	**.507** (.008)
WBC	.612 (.006)	.331 (.010)	.217 (.012)	.699 (.005)	**.749** (.011)	.608 (.007)	.622 (.003)
Plt	.698 (.005)	.706 (.007)	.531 (.011)	.713 (.005)	**.763** (.005)	.703 (.007)	.714 (.004)
Death	.241 (.010)	.209 (.021)	.151 (.020)	.144 (.008)	**.250** (.019)	.221 (.016)	.224 (.010)
Readm.	.451 (.003)	.524 (.005)	.491 (.006)	.481 (.003)	.481 (.004)	.544 (.005)	**.548** (.004)
ED	.359 (.002)	.393 (.005)	.363 (.007)	.408 (.005)	**.439** (.005)	.401 (.007)	.420 (.004)
macro	.464 (.002)	.440 (.004)	.360 (.005)	.495 (.002)	**.531** (.004)	.496 (.003)	.511 (.003)
gen. tokens / traj	5.8	23.3	154.7	95.8	167.3	5.2	5.2
ms / traj	22.2	95.1	592.4	379.8	623.9	21.5	21.5

Table A.8: Per-outcome AUPRC (↑) and generation cost per trajectory at 28 days, 30 days, 42 days and 6 months, EHR-SHOT.

	month	week	day	event stream	*ETHOS*	**MuRGEHR**	**MuRGEHR**
# of gen. traj	100	100	100	100	100	100	1098
28 days							
Creat.	.746 (.003)	.785 (.002)	.807 (.002)	.790 (.002)	.814 (.003)	.807 (.004)	**.816** (.001)
Hct	.573 (.002)	.637 (.002)	.681 (.002)	.676 (.002)	.688 (.003)	.675 (.002)	**.689** (.001)
Hgb	.612 (.003)	.674 (.002)	.718 (.002)	.717 (.002)	.721 (.003)	.726 (.002)	**.739** (.001)
WBC	.633 (.002)	.660 (.002)	.689 (.001)	.669 (.002)	.693 (.002)	.690 (.002)	**.698** (.001)
Plt	.646 (.002)	.665 (.002)	.715 (.002)	.709 (.002)	.723 (.002)	.721 (.002)	**.731** (.001)
Death	.140 (.004)	.160 (.004)	.141 (.004)	**.218** (.006)	.215 (.007)	.169 (.004)	.184 (.004)
Readm.	.617 (.001)	.648 (.001)	.670 (.001)	.673 (.001)	.680 (.003)	.684 (.001)	**.695** (.001)
macro	.567 (.001)	.604 (.001)	.631 (.001)	.636 (.001)	.648 (.002)	.639 (.001)	**.650** (.001)
gen. tokens / traj.	1	3.9	27.5	46.2	187.9	1	1
ms / traj.	2.4	7.9	50.1	77.0	570.1	3.2	3.2
30 days							
Creat.	–	–	.806 (.002)	.788 (.002)	.815 (.003)	.808 (.003)	**.819** (.001)
Hct	–	–	.678 (.002)	.675 (.002)	.686 (.003)	.674 (.002)	**.688** (.001)
Hgb	–	–	.716 (.002)	.715 (.002)	.720 (.003)	.726 (.002)	**.737** (.001)
WBC	–	–	.693 (.001)	.674 (.002)	.696 (.002)	.692 (.002)	**.701** (.001)
Plt	–	–	.711 (.002)	.705 (.002)	.719 (.002)	.712 (.002)	**.727** (.001)
Death	–	–	.144 (.005)	**.221** (.006)	.220 (.006)	.177 (.004)	.187 (.002)
Readm.	–	–	.676 (.001)	.678 (.001)	.687 (.002)	.688 (.001)	**.699** (.000)
macro	–	–	.632 (.001)	.637 (.001)	.649 (.002)	.639 (.001)	**.651** (.000)
gen. tokens / traj.	–	–	29.4	46.2	187.9	4.9	4.9
ms / traj.	–	–	53.8	77.0	570.1	15.9	15.9
42 days							
Creat.	–	.803 (.002)	.813 (.002)	.800 (.002)	.825 (.003)	.823 (.003)	**.833** (.001)
Hct	–	.632 (.002)	.671 (.002)	.668 (.002)	.679 (.003)	.675 (.002)	**.682** (.001)
Hgb	–	.671 (.002)	.709 (.002)	.708 (.002)	.713 (.003)	.719 (.002)	**.729** (.001)
WBC	–	.685 (.002)	.711 (.001)	.690 (.001)	.709 (.002)	.706 (.002)	**.716** (.001)
Plt	–	.663 (.002)	.706 (.002)	.702 (.002)	.720 (.002)	.709 (.002)	**.723** (.001)
Death	–	.187 (.004)	.161 (.005)	.224 (.006)	**.234** (.006)	.189 (.004)	.201 (.002)
Readm.	–	.695 (.001)	.707 (.001)	.707 (.001)	.719 (.002)	.718 (.001)	**.726** (.000)
macro	–	.619 (.001)	.640 (.001)	.643 (.001)	.657 (.002)	.648 (.001)	**.659** (.000)
gen. tokens / traj.	–	5.7	40.9	46.2	187.9	3.9	3.9
ms / traj.	–	11.4	74.8	77.0	570.1	12.6	12.6
6 months							
Creat.	.812 (.001)	.822 (.002)	.827 (.001)	.822 (.002)	.832 (.002)	.832 (.001)	**.840** (.000)
Hct	.592 (.002)	.618 (.002)	.644 (.001)	.645 (.001)	.648 (.002)	.641 (.002)	**.656** (.001)
Hgb	.638 (.002)	.665 (.002)	.687 (.001)	.693 (.001)	.691 (.002)	.694 (.002)	**.703** (.001)
WBC	.744 (.001)	.748 (.001)	.765 (.001)	.762 (.001)	.761 (.001)	.761 (.002)	**.771** (.001)
Plt	.652 (.002)	.655 (.002)	.678 (.001)	.681 (.002)	.690 (.002)	.683 (.002)	**.694** (.001)
Death	.242 (.003)	.255 (.003)	.239 (.003)	**.293** (.003)	.274 (.004)	.262 (.004)	.267 (.002)
Readm.	.784 (.001)	.796 (.001)	.797 (.001)	.801 (.001)	.806 (.001)	.803 (.001)	**.811** (.000)
macro	.638 (.001)	.651 (.001)	.662 (.001)	.671 (.001)	.672 (.001)	.668 (.001)	**.677** (.000)
gen. tokens / traj.	5.2	21.4	147.9	46.2	187.9	5.4	5.4
ms / traj.	11.1	42.7	295.9	77.0	570.1	18.8	18.8

Table A.9: Per-outcome AUPRC at 28 days, 30 days, 42 days and 6 months, Hospital A.

A.4 ECE

	month	week	day	event stream	*ETHOS*	MuRGEHR	MuRGEHR
# of gen. traj	100	100	100	100	100	100	643
28 days							
Creat.	**.0040** (.0006)	.0206 (.0011)	.0294 (.0006)	.0190 (.0008)	.0331 (.0025)	.0081 (.0007)	.0055 (.0006)
Hct	**.0057** (.0011)	.0381 (.0008)	.0440 (.0004)	.0126 (.0014)	.0366 (.0037)	.0075 (.0024)	.0072 (.0027)
Hgb	**.0036** (.0018)	.0348 (.0013)	.0427 (.0006)	.0121 (.0029)	.0405 (.0026)	.0065 (.0026)	.0046 (.0008)
WBC	.0154 (.0043)	.0588 (.0009)	.0495 (.0009)	.0339 (.0032)	.0700 (.0053)	.0082 (.0044)	**.0067** (.0050)
Plt	.0056 (.0008)	.0204 (.0005)	.0169 (.0006)	.0097 (.0010)	.0154 (.0012)	.0060 (.0015)	**.0040** (.0024)
Death	.0114 (.0011)	**.0025** (.0030)	.0110 (.0004)	.0134 (.0001)	.0066 (.0026)	.0086 (.0020)	.0079 (.0004)
Readm.	.0255 (.0068)	.0513 (.0029)	.0653 (.0006)	**.0176** (.0066)	.0730 (.0086)	.0279 (.0083)	.0298 (.0087)
ICU	.0109 (.0031)	.0183 (.0012)	.0181 (.0021)	.0117 (.0024)	.0139 (.0028)	.0091 (.0023)	**.0054** (.0038)
ED	.0246 (.0062)	.0190 (.0056)	.0228 (.0050)	**.0146** (.0057)	.0523 (.0079)	.0252 (.0105)	.0197 (.0073)
macro	.0118 (.0011)	.0293 (.0008)	.0333 (.0006)	.0161 (.0013)	.0379 (.0015)	.0119 (.0017)	**.0101** (.0016)
gen. tokens / traj	2.0	4.8	26.9	39.6	120.1	1.0	1.0
ms / traj	4.8	16.6	119.3	167.4	414.5	5.1	5.1
30 days							
Creat.	–	–	.0307 (.0005)	.0184 (.0008)	.0321 (.0016)	.0138 (.0006)	**.0116** (.0002)
Hct	–	–	.0469 (.0004)	**.0125** (.0016)	.0367 (.0030)	.0153 (.0007)	.0140 (.0001)
Hgb	–	–	.0450 (.0005)	**.0120** (.0031)	.0402 (.0027)	.0139 (.0008)	.0124 (.0003)
WBC	–	–	.0505 (.0010)	.0304 (.0030)	.0664 (.0042)	.0169 (.0016)	**.0157** (.0001)
Plt	–	–	.0175 (.0007)	.0089 (.0006)	.0147 (.0008)	.0061 (.0007)	**.0035** (.0020)
Death	–	–	.0111 (.0002)	.0143 (.0001)	**.0070** (.0026)	.0092 (.0013)	.0092 (.0005)
Readm.	–	–	.0667 (.0004)	.0215 (.0075)	.0776 (.0086)	.0223 (.0080)	**.0196** (.0138)
ICU	–	–	.0189 (.0013)	.0114 (.0027)	.0137 (.0029)	.0067 (.0046)	**.0060** (.0060)
ED	–	–	.0199 (.0052)	**.0160** (.0071)	.0567 (.0079)	.0258 (.0041)	.0260 (.0060)
macro	–	–	.0341 (.0007)	.0162 (.0013)	.0384 (.0015)	.0145 (.0012)	**.0131** (.0018)
gen. tokens / traj	–	–	28.7	39.6	120.1	4.9	4.9
ms / traj	–	–	169.7	167.4	414.5	26.8	26.8
42 days							
Creat.	–	.0233 (.0009)	.0347 (.0007)	.0147 (.0011)	.0278 (.0019)	.0132 (.0006)	**.0114** (.0008)
Hct	–	.0569 (.0007)	.0584 (.0005)	**.0114** (.0017)	.0326 (.0027)	.0177 (.0008)	.0157 (.0001)
Hgb	–	.0505 (.0006)	.0553 (.0009)	**.0110** (.0040)	.0358 (.0035)	.0148 (.0009)	.0128 (.0004)
WBC	–	.0780 (.0009)	.0587 (.0012)	.0232 (.0036)	.0540 (.0063)	**.0215** (.0020)	.0216 (.0001)
Plt	–	.0270 (.0008)	.0216 (.0006)	.0072 (.0012)	.0115 (.0016)	.0068 (.0014)	**.0049** (.0009)
Death	–	**.0039** (.0029)	.0117 (.0008)	.0198 (.0001)	.0112 (.0027)	.0109 (.0021)	.0113 (.0023)
Readm.	–	.0819 (.0034)	.0805 (.0004)	.0345 (.0011)	.0951 (.0088)	.0341 (.0084)	**.0309** (.0088)
ICU	–	.0312 (.0009)	.0268 (.0015)	**.0100** (.0030)	.0118 (.0034)	.0131 (.0031)	.0126 (.0001)
ED	–	**.0188** (.0077)	.0258 (.0054)	.0217 (.0025)	.0680 (.0082)	.0458 (.0007)	.0451 (.0001)
macro	–	.0413 (.0010)	.0415 (.0007)	**.0170** (.0008)	.0387 (.0017)	.0198 (.0011)	.0185 (.0010)
gen. tokens / traj	–	6.5	39.3	39.6	120.1	3.9	3.9
ms / traj	–	36.8	187.9	167.4	414.5	19.9	19.9
6 months							
Creat.	.0161 (.0013)	.0304 (.0007)	.0554 (.0008)	**.0119** (.0029)	.0209 (.0020)	.0463 (.0004)	.0457 (.0001)
Hct	.0611 (.0004)	.1331 (.0007)	.1156 (.0006)	**.0094** (.0025)	.0232 (.0032)	.0945 (.0005)	.0941 (.0001)
Hgb	.0599 (.0006)	.1233 (.0006)	.1110 (.0003)	**.0138** (.0025)	.0299 (.0035)	.0860 (.0006)	.0864 (.0001)
WBC	.0595 (.0014)	.1451 (.0003)	.0846 (.0015)	**.0204** (.0045)	.0359 (.0040)	.1046 (.0004)	.1053 (.0002)
Plt	.0152 (.0010)	.0785 (.0007)	.0486 (.0014)	**.0103** (.0026)	.0143 (.0019)	.0488 (.0005)	.0483 (.0001)
Death	.0233 (.0006)	.0569 (.0005)	.0152 (.0036)	.0426 (.0002)	.0317 (.0034)	**.0120** (.0041)	.0155 (.0060)
Readm.	.0558 (.0050)	.1238 (.0005)	.1005 (.0049)	**.0321** (.0005)	.1121 (.0088)	.0684 (.0005)	.0683 (.0002)
ICU	**.0142** (.0076)	.0990 (.0004)	.0894 (.0008)	.0211 (.0047)	.0195 (.0066)	.0491 (.0004)	.0491 (.0001)
ED	.0324 (.0047)	.0425 (.0056)	.0564 (.0047)	.0305 (.0004)	.0930 (.0081)	.0202 (.0064)	**.0179** (.0115)
macro	.0375 (.0011)	.0925 (.0006)	.0752 (.0009)	**.0213** (.0010)	.0423 (.0025)	.0589 (.0008)	.0590 (.0014)
gen. tokens / traj	5.5	17.6	129.9	39.6	120.1	5.0	5.0
ms / traj	23.0	113.0	744.4	167.4	414.5	36.4	36.4

Table A.10: Per-outcome ECE (↓, lower is better) and generation cost per trajectory at 28 days, 30 days, 42 days and 6 months, MIMIC.

	month	week	day	event stream	*ETHOS*	MuRGEHR	MuRGEHR
# of gen. traj	100	100	100	100	100	100	691
28 days							
Creat.	.0240 (.0041)	.0353 (.0005)	.1778 (.0007)	.0212 (.0024)	**.0161** (.0031)	.0210 (.0011)	.0204 (.0024)
Hct	.0860 (.0045)	.0402 (.0040)	.1227 (.0017)	.0390 (.0030)	.0428 (.0057)	.0301 (.0022)	**.0288** (.0065)
Hgb	.0692 (.0052)	.0398 (.0028)	.1294 (.0009)	.0415 (.0026)	.0396 (.0044)	**.0328** (.0021)	.0331 (.0029)
WBC	**.0171** (.0029)	.1180 (.0007)	.2515 (.0008)	.0287 (.0047)	.0361 (.0018)	.0493 (.0006)	.0495 (.0002)
Plt	.0441 (.0046)	.0506 (.0016)	.2087 (.0007)	**.0264** (.0051)	.0355 (.0025)	.0627 (.0084)	.0643 (.0003)
Death	.0088 (.0029)	.0121 (.0050)	.0121 (.0010)	.0103 (.0022)	.0112 (.0015)	.0066 (.0039)	**.0058** (.0033)
Readm.	.1806 (.0057)	.0735 (.0137)	.1903 (.0017)	.1198 (.0061)	.1124 (.0060)	**.0450** (.0065)	.0473 (.0048)
ED	.1216 (.0046)	.0341 (.0057)	.0362 (.0103)	.0510 (.0077)	.0381 (.0098)	.0392 (.0181)	**.0246** (.0209)
macro	.0689 (.0018)	.0504 (.0021)	.1411 (.0014)	.0422 (.0016)	.0415 (.0018)	.0358 (.0027)	**.0342** (.0029)
gen. tokens / traj	1.0	4.0	27.4	95.8	167.3	1.0	1.0
ms / traj	5.5	15.0	106.8	379.8	623.9	4.7	4.7
30 days							
Creat.	–	–	.1826 (.0007)	.0217 (.0026)	**.0170** (.0042)	.0223 (.0011)	.0213 (.0002)
Hct	–	–	.1237 (.0023)	.0388 (.0029)	.0417 (.0046)	.0287 (.0038)	**.0271** (.0003)
Hgb	–	–	.1324 (.0016)	.0411 (.0035)	.0407 (.0039)	**.0309** (.0016)	.0320 (.0030)
WBC	–	–	.2603 (.0008)	**.0304** (.0068)	.0360 (.0018)	.0487 (.0006)	.0492 (.0002)
Plt	–	–	.2137 (.0007)	**.0283** (.0053)	.0373 (.0027)	.0618 (.0067)	.0630 (.0003)
Death	–	–	.0123 (.0002)	.0105 (.0021)	.0114 (.0016)	**.0065** (.0059)	.0072 (.0072)
Readm.	–	–	.1924 (.0044)	.1206 (.0066)	.1107 (.0085)	.0539 (.0052)	**.0464** (.0099)
ED	–	–	.0351 (.0124)	.0484 (.0075)	.0403 (.0086)	.0294 (.0172)	**.0229** (.0229)
macro	–	–	.1441 (.0017)	.0425 (.0018)	.0419 (.0019)	.0353 (.0024)	**.0336** (.0032)
gen. tokens / traj	–	–	29.4	95.8	167.3	4.7	4.7
ms / traj	–	–	115.4	379.8	623.9	20.2	20.2
42 days							
Creat.	–	.0461 (.0005)	.2077 (.0007)	.0239 (.0013)	**.0154** (.0019)	.0301 (.0011)	.0296 (.0002)
Hct	–	.0484 (.0024)	.1376 (.0023)	.0444 (.0022)	.0449 (.0048)	.0385 (.0089)	**.0347** (.0118)
Hgb	–	.0463 (.0029)	.1454 (.0012)	.0449 (.0029)	.0442 (.0051)	.0340 (.0012)	**.0332** (.0003)
WBC	–	.1574 (.0008)	.3039 (.0008)	**.0252** (.0012)	.0315 (.0021)	.0552 (.0010)	.0550 (.0002)
Plt	–	.0698 (.0026)	.2358 (.0008)	**.0270** (.0088)	.0376 (.0033)	.0668 (.0067)	.0696 (.0003)
Death	–	.0162 (.0014)	.0167 (.0021)	.0119 (.0056)	.0144 (.0009)	**.0087** (.0072)	.0122 (.0064)
Readm.	–	.0876 (.0088)	.2009 (.0084)	.1107 (.0070)	.0977 (.0138)	.0552 (.0065)	**.0497** (.0109)
ED	–	.0359 (.0169)	.0354 (.0127)	.0603 (.0155)	.0485 (.0098)	.0381 (.0224)	**.0313** (.0117)
macro	–	.0635 (.0024)	.1604 (.0018)	.0435 (.0026)	.0418 (.0026)	.0408 (.0034)	**.0394** (.0026)
gen. tokens / traj	–	6.0	41.0	95.8	167.3	3.8	3.8
ms / traj	–	22.6	200.3	379.8	623.9	22.8	22.8
6 months							
Creat.	.0138 (.0003)	.0907 (.0007)	.2755 (.0007)	.0202 (.0011)	**.0092** (.0036)	.0832 (.0007)	.0819 (.0003)
Hct	.1312 (.0045)	.0850 (.0026)	.1524 (.0031)	**.0403** (.0092)	.0443 (.0088)	.0950 (.0008)	.0968 (.0003)
Hgb	.1084 (.0063)	.0772 (.0037)	.1627 (.0007)	**.0368** (.0079)	.0543 (.0090)	.0834 (.0008)	.0841 (.0003)
WBC	.0430 (.0003)	.2985 (.0011)	.4816 (.0010)	.0284 (.0040)	**.0251** (.0039)	.1685 (.0009)	.1682 (.0004)
Plt	.0550 (.0004)	.1427 (.0008)	.2980 (.0008)	**.0277** (.0069)	.0351 (.0039)	.1540 (.0009)	.1552 (.0003)
Death	.0437 (.0003)	.0464 (.0003)	.0479 (.0014)	.0430 (.0003)	.0554 (.0013)	**.0267** (.0058)	.0294 (.0051)
Readm.	.2203 (.0110)	.1555 (.0098)	.2352 (.0010)	.1003 (.0077)	**.0742** (.0128)	.0882 (.0188)	.0893 (.0151)
ED	.1752 (.0093)	.0714 (.0089)	.0950 (.0087)	**.0608** (.0202)	.0754 (.0060)	.0881 (.0092)	.0848 (.0004)
macro	.0988 (.0021)	.1209 (.0018)	.2185 (.0013)	**.0447** (.0034)	.0466 (.0027)	.0984 (.0027)	.0987 (.0020)
gen. tokens / traj	5.8	23.3	154.7	95.8	167.3	5.2	5.2
ms / traj	22.2	95.1	592.4	379.8	623.9	21.5	21.5

Table A.11: Per-outcome ECE (↓, lower is better) and generation cost per trajectory at 28 days, 30 days, 42 days and 6 months, EHR-SHOT.

	month	week	day	event stream	*ETHOS*	**MuRGEHR**	**MuRGEHR**
# of gen. traj	100	100	100	100	100	100	1098
28 days							
Creat.	**.007** (.002)	.010 (.001)	.011 (.001)	.011 (.001)	.014 (.002)	.015 (.002)	.012 (.001)
Hct	.012 (.003)	.009 (.003)	.011 (.003)	**.008** (.002)	.032 (.004)	.013 (.003)	.011 (.003)
Hgb	.013 (.002)	**.008** (.003)	.017 (.002)	**.008** (.001)	.029 (.004)	.015 (.003)	.015 (.002)
WBC	**.011** (.004)	.041 (.000)	.014 (.004)	.041 (.001)	.024 (.004)	.026 (.003)	.025 (.003)
Plt	**.005** (.002)	.006 (.002)	.007 (.001)	.007 (.003)	.011 (.002)	.011 (.002)	.011 (.002)
Death	.005 (.002)	.005 (.002)	.005 (.003)	.007 (.002)	.006 (.003)	.007 (.002)	**.003** (.002)
Readm.	**.023** (.004)	.039 (.001)	.086 (.000)	.044 (.007)	.024 (.007)	.048 (.000)	.049 (.003)
macro	**.011** (.001)	.017 (.001)	.022 (.001)	.018 (.001)	.020 (.002)	.019 (.001)	.018 (.001)
gen. tokens / traj.	1	3.9	27.5	46.2	187.9	1	1
ms / traj.	2.4	7.9	50.1	77.0	570.1	3.2	3.2
30 days							
Creat.	–	–	**.012** (.001)	**.012** (.001)	.013 (.002)	.015 (.002)	**.012** (.002)
Hct	–	–	.012 (.002)	**.008** (.003)	.033 (.004)	.013 (.003)	.011 (.003)
Hgb	–	–	.017 (.003)	**.008** (.001)	.029 (.004)	.016 (.002)	.014 (.003)
WBC	–	–	**.013** (.003)	.039 (.001)	.024 (.005)	.027 (.002)	.026 (.002)
Plt	–	–	.008 (.001)	**.007** (.003)	.010 (.002)	.013 (.002)	.011 (.003)
Death	–	–	.005 (.002)	.008 (.002)	.005 (.002)	.005 (.001)	**.003** (.001)
Readm.	–	–	.084 (.000)	.042 (.005)	**.023** (.009)	.043 (.000)	.044 (.002)
macro	–	–	.021 (.001)	.018 (.001)	.020 (.002)	.019 (.001)	**.017** (.001)
gen. tokens / traj.	–	–	29.4	46.2	187.9	4.9	4.9
ms / traj.	–	–	53.8	77.0	570.1	15.9	15.9
42 days							
Creat.	–	**.010** (.000)	.013 (.001)	.012 (.001)	.012 (.002)	.014 (.002)	.011 (.003)
Hct	–	.011 (.002)	.013 (.002)	**.010** (.003)	.032 (.004)	.013 (.002)	.011 (.002)
Hgb	–	.009 (.003)	.016 (.001)	**.008** (.003)	.027 (.004)	.015 (.003)	.013 (.004)
WBC	–	.054 (.000)	**.014** (.001)	.035 (.001)	.020 (.004)	.035 (.001)	.034 (.000)
Plt	–	.010 (.003)	.010 (.001)	**.008** (.003)	.009 (.002)	.014 (.002)	.012 (.002)
Death	–	**.003** (.003)	.006 (.003)	.011 (.000)	.006 (.002)	.006 (.001)	.004 (.000)
Readm.	–	.042 (.001)	.084 (.000)	.032 (.004)	**.029** (.007)	.040 (.000)	.040 (.000)
macro	–	.020 (.001)	.022 (.001)	**.017** (.001)	.019 (.002)	.020 (.001)	.018 (.001)
gen. tokens / traj.	–	5.7	40.9	46.2	187.9	3.9	3.9
ms / traj.	–	11.4	74.8	77.0	570.1	12.6	12.6
6 months							
Creat.	**.009** (.001)	**.009** (.002)	.014 (.001)	.020 (.001)	.011 (.003)	.028 (.002)	.028 (.002)
Hct	.027 (.001)	.036 (.002)	.027 (.003)	.025 (.002)	**.024** (.004)	.061 (.001)	.061 (.000)
Hgb	.021 (.001)	.030 (.002)	.029 (.002)	.026 (.002)	**.015** (.005)	.051 (.000)	.051 (.000)
WBC	.043 (.001)	.112 (.000)	.027 (.001)	.027 (.002)	**.021** (.005)	.125 (.000)	.126 (.000)
Plt	**.009** (.002)	.025 (.002)	.014 (.002)	.022 (.000)	.014 (.004)	.040 (.002)	.041 (.003)
Death	.014 (.005)	.011 (.003)	.026 (.002)	.031 (.000)	**.009** (.004)	.030 (.001)	.030 (.000)
Readm.	.043 (.003)	.053 (.001)	.069 (.000)	**.024** (.007)	.027 (.007)	.100 (.000)	.100 (.000)
macro	.024 (.001)	.039 (.001)	.029 (.001)	.025 (.001)	**.017** (.002)	.062 (.000)	.062 (.001)
gen. tokens / traj.	5.2	21.4	147.9	46.2	187.9	5.4	5.4
ms / traj.	11.1	42.7	295.9	77.0	570.1	18.8	18.8

Table A.12: Per-outcome expected calibration error (ECE) at 28 days, 30 days, 42 days and 6 months, Hospital A.

B DATA PROCESSING

Sources and splits. We use MIMIC-IV (Johnson et al., 2024), EHR-SHOT (Wornow et al., 2023), and Hospital A and report metrics in Table A.1 and A.2. We split by patient IDs for MIMIC and Hospital A, and we use the provided patient splits for EHRSHOT. We uses an aggressive subset of the original codes to make generation tractable for the fine-grained baselines. This filtering only leaves codes for evaluated outcomes, trajectory boundaries (admissions and discharges), leukocyte, platelet, hemoglobin, creatinine, and hematocrit labs, death, and birth Oufattole et al.. We define the abnormal labs to be greater than 2.3 mg/dL for creatinine, less than 24.6% for hematocrit, less than 8.0 g/dL for hemoglobin, less than 5.1 K/uL for leukocyte (WBC), and less than 86.0 K/uL for platelet Oufattole et al.. We use age as a feature for MIMIC and Hospital A, and use demographic information, such as gender, race, and ethnicity.

We add an end-of-sequence token to the codeset and do not fill with end of sequence tokens, it is instead padded with empty tokens. These padded zeros are not included in the loss. During generation we truncate at the first eos. We have an eos per resolution.

C TRAINING DETAILS

C.1 OUR MODELS

We train all of our models using AdamW at a peak learning rate of 3e-4, with 5,000 steps of linear warmup and cosine decay to 1e-5. On our experiments, the schedule leads to about 2.5x faster convergence. We use early stopping on the validation loss with an absolute improvement to the validation loss (1e-4) and a batch size of 256. We set the maximum iterations to 100k.

C.2 ETHOS

We train Ethos following https://github.com/Oufattole/ethos-eic. We use the same architecture described in the original implementation of 6 layers, 12 heads, and 768 width. We change the context size (number of tokens) to match our models.

C.3 INFERENCE

For the latency and token usage per trajectory, we use 50 warmup iterations and average over the following 100 iterations. MIMIC and EHRShHT were run on nvidia V100s while Hospital A was run on nvidia A100s.

For the baseline single resolution autoregressive models and the implemented event stream, we use KV Caching during rollout.