Does Temporal Encoding Matter? Evidence from GRU-based Clinical Foundation Models

Anonymous Authors[1]

Abstract

Clinical foundation models for structured electronic health records handle temporal irregularity in varied ways, yet the impact of time encoding strategy on downstream performance remains poorly understood. While access to longer patient histories has been shown to improve predictive performance, there is no consensus on the consistent benefit of explicit time encoding in clinical foundation models. We train gated recurrent unit (GRU) models under five temporal encoding conditions, namely no encoding, discrete time tokens, learned projection, fixed sinusoidal encoding, and learned sinusoidal encoding on MIMIC-IV data, and evaluate downstream performance across six clinical prediction tasks using randomly sampled admission points spanning the full patient timeline. We find that the temporal encoding strategy has no significant impact on downstream performance, even for the patients with the longest sequence lengths and multiple historical visits. This implies that the ordering of data either sufficiently encodes temporal dependencies or that the dataset lacks the temporal diversity to realise the potential improvements from a deeper understanding of time. Our findings caution against assuming that richer time representations improve clinical prediction, and motivate evaluation on datasets with greater temporal heterogeneity.

1. Introduction

Clinical foundation models trained with self-supervised objectives on electronic healthcare record (EHR) data have shown versatility in their performance on a number of downstream prediction tasks, such as hospital readmission and mortality prediction (Steinberg et al., 2021; Renc et al., 2024; Shmatko et al., 2025; Kraljevic et al., 2024). These models are trained on structured medical event sequences including diagnoses, medications, procedures, and laboratory results, embedding clinical information into patient representations that support downstream predictive tasks.

The best performing variants adapt language model architectures to medical event sequences, making use of the sequential and discrete nature of structured medical events. Unlike natural language, however, medical events occur at highly irregular intervals spanning minutes to years, the length of which alters the clinical meaning of those events. To address this, early recurrent approaches incorporated time decays based on elapsed time to down-weight old information (Che et al., 2018), while a more recent GRU-based model, CLMBR, concatenated multidimensional temporal features onto input feature vectors (Guo et al., 2023). Similarly, transformer-based architectures have experimented with discrete time gap tokens (Renc et al., 2024), sinusoidal elapsed-time embeddings (Shmatko et al., 2025), and continuous value embeddings (Tipirneni & Reddy, 2022). The incorporation of time in all of these models is motivated by the assumed information content of the irregular intervals.

Despite the proliferation of these methods, there is no consensus on the most effective time condition or whether explicit time encoding is necessary at all. While transformers with extended context windows have shown improvements in downstream predictive tasks (Wornow et al., 2025), recent evidence suggests that explicit temporal tokenisation may be redundant for transformer-based performance on mortality and readmission predictions (Attrach et al., 2025). It remains unclear whether this redundancy is a feature of input sequence length, attention-based approaches, or is inherent to the specific evaluation tasks on EHR data. To disentangle these factors, we employ a GRU architecture, which has been shown to perform within a narrow margin of transformers on clinical prediction tasks (Tipirneni & Reddy, 2022). This allows us to test whether the null result generalises beyond attention-based models, and to systematically vary input sequence length across the full patient timeline, something transformers are poorly suited to given their practical context window constraints (Vaswani et al., 2017).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Methods

2.1. Data and Pre-training

Hospital electronic health record data from MIMIC-IV (Johnson et al., 2024) were used for this study. Data were processed into Medical Event Data Standard (MEDS) (McDermott et al., 2025) using a modified version of the MEDS Transform pipeline (Vector Institute, 2024).

All models were pre-trained on a next-event prediction objective, structured as a multi-label binary classification task over the entire medical event vocabulary. Given sequences of events up to a randomly sampled anchor point within a patient's record, the models were tasked with predicting the events appearing in the next 24-hour window. During training, five evaluation anchor points were randomly sampled per patient per epoch.

Medical codes at each time step were embedded via a learned embedding layer of dimension $d = 128$ and mean-pooled across codes to produce a single event vector per hour, as in (Tipirneni & Reddy, 2022) and padding tokens were excluded from the pool. This event vector was then optionally augmented with a temporal encoding before being passed to the recurrent encoder. Patients with fewer than 5 hourly time steps were removed from training.

2.2. Model Architecture

The GRU had a hidden dimension of 256, two layers, and dropout of 0.2 between layers. All models were trained for 10 epochs using the AdamW optimiser and a cosine annealing learning rate schedule. The initial learning rate and evaluation window size were selected via 20 trials of hyperparameter tuning (details shown in A.1), and the best epoch's parameters according to validation loss were retained for each parameter set. The optimal hyperparameters were then chosen according to the best precision@10 and recall@10 performance on the held-out test set (20% of the data).

2.3. Temporal Encoding Strategies

We compared four strategies for explicitly incorporating time into the event representations against a baseline model without time encoding. In all cases, the core architecture was held constant, with only the time encoding module changing.

Discrete TIME tokens: Taking inspiration from the ETHOS framework (Renc et al., 2024), we encoded temporal gaps directly as events in the patient event stream by inserting discrete `TIME//N` tokens at inter-event gaps. The continuous inter-event duration was quantised into one of 10 variable-width buckets (Appendix 2).

Sinusoidal encoding: Building on the Delphi architecture (Shmatko et al., 2025), we evaluated sinusoidal time embeddings, extending their work by encoding two temporal features per time-step: elapsed hours since the first recorded event (t_{elapsed}) and hours since the previous event (t_δ). These times were mapped to a d-dimensional embedding using a frequency bank initialised geometrically between $2\pi/T_{\max}$ and $2\pi/T_{\min}$ ($T_{\min} = 1$h, $T_{\max} = 87{,}600$h). For a given time t, the d-dimensional embedding $\mathbf{s}$(t) was constructed as:

$$\mathbf{s}(t) = \big[\sin(\omega_1 t),\ \ldots,\ \sin(\omega_{d/2}\, t),\ \cos(\omega_1 t),\ \ldots,\ \cos(\omega_{d/2}\, t)\big] \in \mathbb{R}^d, \tag{1}$$

Embeddings from both features were summed into the token embedding at each time step before being passed to the GRU. We evaluated two variants: *fixed* frequencies, which remained constant throughout training, and *learnable* frequencies following Xu et al. (2019), where the log-frequencies $\mathbf{f} \in \mathbb{R}^{d/2}$, from which $\boldsymbol{\omega} = \exp(\mathbf{f})$, were optimised jointly with the model, allowing it to discover clinically relevant timescales such as circadian ($\sim$24h) or weekly ($\sim$168h) rhythms. In the learnable variant, frequencies were constrained to remain strictly positive by storing them in log-space and applying a softplus transformation at runtime, preventing degenerate zero or negative frequencies during optimisation.

Learned projection encodings: The log-scaled time features were combined $\boldsymbol{\tau} = [\log(1 + t_{\text{elapsed}}), \log(1 + t_\delta)]$ and projected into the embedding space via a learned linear layer $W_\tau \in \mathbb{R}^{d\times 2}$:

$$\mathbf{p}(\boldsymbol{\tau}) = W_\tau\, \boldsymbol{\tau}. \tag{2}$$

The resulting projection was added element-wise to the medical event vector.

2.4. Clinical Evaluation

Next-event prediction: Models were evaluated based on their ability to predict a set of clinical codes occurring within a 24-hour window of a randomly sampled anchor point. We report Recall@k and Precision@k, where Recall@k measures the proportion of events in the 24-hour window retrieved in the top k predictions and Precision@k measures the fraction of the top k predictions that occurred in the window.

Downstream clinical tasks: Following the approach of CLIMBR (Steinberg et al., 2021), the final hidden state from the GRU was extracted as a frozen patient representation, and was used as the input to logistic regression (LR) and random forest (RF) models to predict six binary clinical tasks. These tasks included predicting three in-visit outcomes, namely mortality, emergency department (ED)

admission, and intensive care unit (ICU) admission, from the first 48 hours of the data from the current visit and all preceding patient data. The other three tasks considered all the data up to the discharge of the randomly chosen visit and involved predicting either the 30-day readmission to the hospital, ED, or ICU.

2.5. Learned Projection Weight Analysis

In order to interpret what the learned projection condition had encoded, we analysed the projections of the linear layer of both the t_{elapsed} and t_δ inputs. We swept each input feature across a range of clinically relevant values while holding the other fixed and computed pairwise cosine similarities between the resulting time vectors.

2.6. Statistical Analysis

Statistical significance was assessed using paired two-sided Wilcoxon signed-rank tests comparing each time-aware condition against the time-unaware baseline, with Holm–Bonferroni correction for multiple comparisons across tasks and metrics.

3. Results

3.1. Next-Event Prediction

Performance on the pre-training objective is summarised in Table 1. The no-time baseline achieved the highest Recall@10 (37.9%) and Precision@10 (38.7%) of all five conditions. This dominance was consistent across all values of k from 1 to 10.

Table 1. Recall@10 and Precision@10 for each model on the next-event prediction task.

Model	Recall@10	Precision@10
Baseline (no explicit time)	**37.9%**	**38.7%**
Learned time embedding	36.3%	37.1%
Sinusoidal (learnt)	33.5%	34.4%
Sinusoidal (fixed)	34.3%	35.3%
Time tokens	32.2%	33.6%

3.2. Exploring the Learned Time Representations

To understand why temporal encoding conferred no downstream benefit, we analysed the projections of both the inter-event gaps and the elapsed times in the learned projection condition (Figure 1). Both components showed a gradual decline in cosine similarity with increasing temporal separation. However, the inter-event gap component declined more steeply, with cosine similarity of representations at the largest separations becoming negative, indicating a stronger discrimination of time-frames.

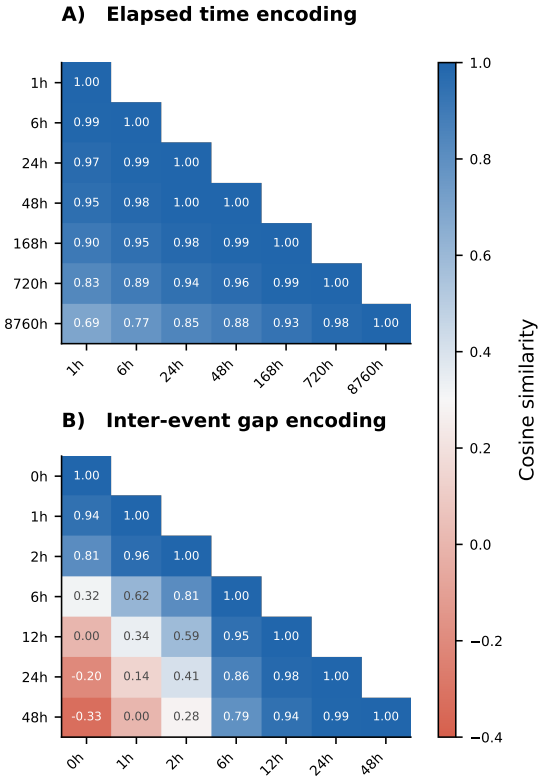

Figure 1. Cosine similarity between time projection vectors at pairs of time points. A) Elapsed time component, evaluated from 1 hour to a year. B) Inter-event gap component, evaluated from 0 to 48 hours

3.3. Downstream Clinical Tasks

Figure 2 shows the AUROC and AUPRC scores for all six clinical tasks stratified by length of sequence. After Holm–Bonferroni correction, all pairwise differences between time-aware conditions and the no-time baseline were statistically indistinguishable. This held across all downstream tasks and across both short and long sequence quartiles.

4. Discussion

Temporal encoding does not show a meaningful improvement on the pre-training objective or downstream clinical prediction performance in this GRU-based EHR foundation model. This result persists across four time encoding conditions in six clinical prediction tasks, two input regimes, and across sequence length quartiles. The finding is consistent with prior null results in transformer models (Attrach et al., 2025), and suggests that time is either not being encoded sufficiently or is not being utilised in the downstream evaluations.

The no-time baseline outperformed all time conditions on the pre-training objective across all values of k, suggesting that added time features may actively degrade performance

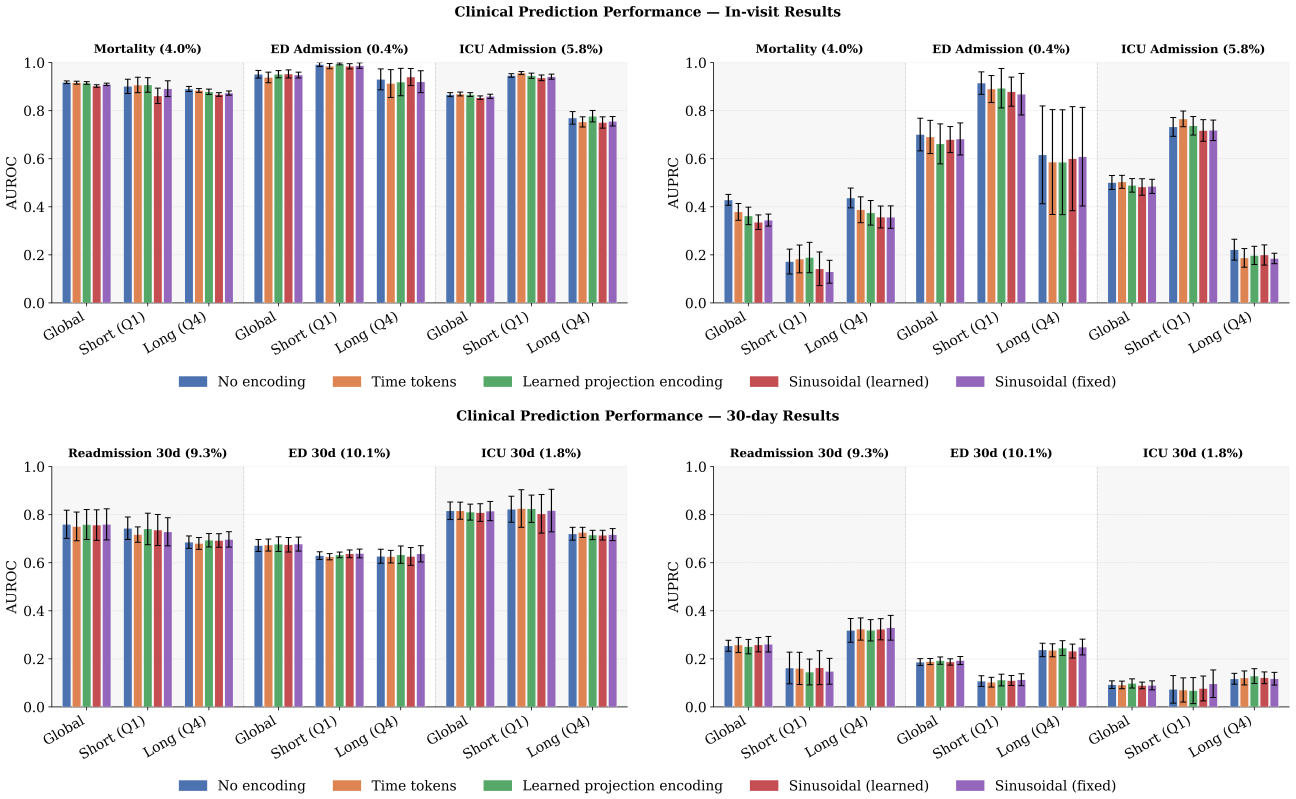

Figure 2. **Downstream Performance of GRU foundation model final embeddings under different time encoding conditions.** Comparison of AUROC and AUPRC (mean across 10 seeds ± 95% confidence interval) across six clinical prediction tasks. All historical events preceding the randomly chosen evaluation visit are included. Prevalences for each task are shown in parentheses following the task name above each collection of Global, Short (first token length quartile) and Long (fourth token length quartile) sequence length results.

on the pre-training task. This may be explained by the dominance in importance of the most recent events in the pre-training task, such that the deeper encoding of historical event timing introduces noise into the hidden state rather than useful context. Given that recent events may dominate the hidden state and mask the utility of temporal encoding, we performed an ablation in which all current-visit data was excluded leaving only pre-admission data (Appendix 5). This found that adding time encoding did not meaningfully improve performance on the clinical prediction tasks, implying that the no-time baseline already captures equally useful medical information from the token order and prevalence alone.

This failure to improve could be explained by the dataset characteristics. With only 2.46 visits on average per individual and most events occurring within the one-hour time bin following the previous event (Appendix A.2), the diversity of temporal information is limited. This is supported by the mechanistic analysis of learned projection weights. Whilst the inter-event gap component produces geometrically distinct representations, with cosine similarity declining systematically with increasing temporal distance (Figure 1), this structure is not exploited in the downstream predictions. Together, these results imply that the dataset does not reward explicit temporal encoding, which likely explains why patients with longer sequence histories showed no differential gain.

The generalisability of this redundancy may be bounded by the nature of the evaluation tasks and the dataset itself. MIMIC-IV is a high-density critical care dataset that likely relies less on long-term temporal patterns than other clinical settings, and the prediction tasks used here are largely determined by short-term clinical trajectories. Establishing whether these findings persist across more temporally heterogeneous datasets and longer-horizon evaluation tasks remains an important open question.

Impact Statement

This paper presents work whose goal is to advance the field of EHR based foundation models by exploring the impact that time encodings can have on prediction performance. By identifying the limitations of current approaches, we hope to motivate future work towards more effective use of temporal information in clinical prediction. This work has potential applications in areas outside the healthcare sector, none of

which, we feel, must be specifically highlighted here.

References

Attrach, R. A., Fani, R., Restrepo, D., Jia, Y., and Schüffler, P. Rethinking Tokenization for Clinical Time Series: When Less is More, December 2025. URL http://arxiv.org/abs/2512.05217. arXiv:2512.05217 [cs].

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1): 6085, April 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-24271-9. URL https://www.nature.com/articles/s41598-018-24271-9.

Guo, L. L., Steinberg, E., Fleming, S. L., Posada, J., Lemmon, J., Pfohl, S. R., Shah, N., Fries, J., and Sung, L. EHR foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports*, 13 (1):3767, March 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-30820-8. URL https://www.nature.com/articles/s41598-023-30820-8.

Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., and Mark, R. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. URL https://doi.org/10.13026/kpb9-mt58. Version 3.1.

Kraljevic, Z., Bean, D., Shek, A., Bendayan, R., Hemingway, H., Yeung, J. A., Deng, A., Balston, A., Ross, J., Idowu, E., Teo, J. T., and Dobson, R. J. B. 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, April 2024. ISSN 2589-7500. doi: 10.1016/S2589-7500(24)00025-6. URL https://www.thelancet.com/journals/landig/article/PIIS2589-7500(24)00025-6/fulltext.

McDermott, M. B. A., Xu, J., Bergamaschi, T. S., Jeong, H., Lee, S. A., Oufattole, N., Rockenschaub, P., Stankevičiūtundefined, K., Steinberg, E., Sun, J., Water, R. P. v. d., Wornow, M., Wu, J., and Wu, Z. Meds: Building models and tools in a reproducible health ai ecosystem. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 6243–6244, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3737608. URL https://doi.org/10.1145/3711896.3737608.

Renc, P., Jia, Y., Samir, A. E., Was, J., Li, Q., Bates, D. W., and Sitek, A. Zero shot health trajectory prediction using transformer. *npj Digital Medicine*, 7(1): 256, September 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01235-0. URL https://www.nature.com/articles/s41746-024-01235-0.

Shmatko, A., Jung, A. W., Gaurav, K., Brunak, S., Mortensen, L. H., Birney, E., Fitzgerald, T., and Gerstung, M. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, November 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09529-3. URL https://www.nature.com/articles/s41586-025-09529-3.

Steinberg, E., Jung, K., Fries, J. A., Corbin, C. K., Pfohl, S. R., and Shah, N. H. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, January 2021. ISSN 1532-0464. doi: 10.1016/j.jbi.2020.103637. URL https://www.sciencedirect.com/science/article/pii/S1532046420302653.

Tipirneni, S. and Reddy, C. K. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Trans. Knowl. Discov. Data*, 16(6), July 2022. ISSN 1556-4681. doi: 10.1145/3516367. URL https://doi.org/10.1145/3516367.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Vector Institute. MEDS transform odyssey branch. https://github.com/VectorInstitute/meds/tree/odyssey, 2024. Accessed: 2026-01-25.

Wornow, M., Bedi, S., Hernandez, M. A. F., Steinberg, E., Fries, J. A., Re, C., Koyejo, S., and Shah, N. H. Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHRs, March 2025. URL http://arxiv.org/abs/2412.16178. arXiv:2412.16178 [cs].

Xu, D., Ruan, C., Korpeoglu, E., Kumar, S., and Achan, K. Self-attention with functional time representation learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

A. Appendix

A.1. Hyperparameter Tuning

Hyperparameter tuning was performed using the Weights and Biases *Sweeps* method. Bayesian optimisation was used to select the next set of hyperparameters to test from a pre-specified range of possible values using validation loss as the performance metric. Twenty hyperparameter combinations were considered for each model. The learning rate range was from 10^{-2} to 10^{-4}, and the size of the prediction window was selected as 24 or 168 hours. Additionally, the weight decay for AdamW was set at 10^{-4}, and the number of evaluation anchors was set at five. For each encoding strategy the best performing model used a window size of 24 hours and the optimal learning rates tended towards the lower end of the learning rate range.

A.2. Distribution of Elapsed Times and Inter-event Time Gaps

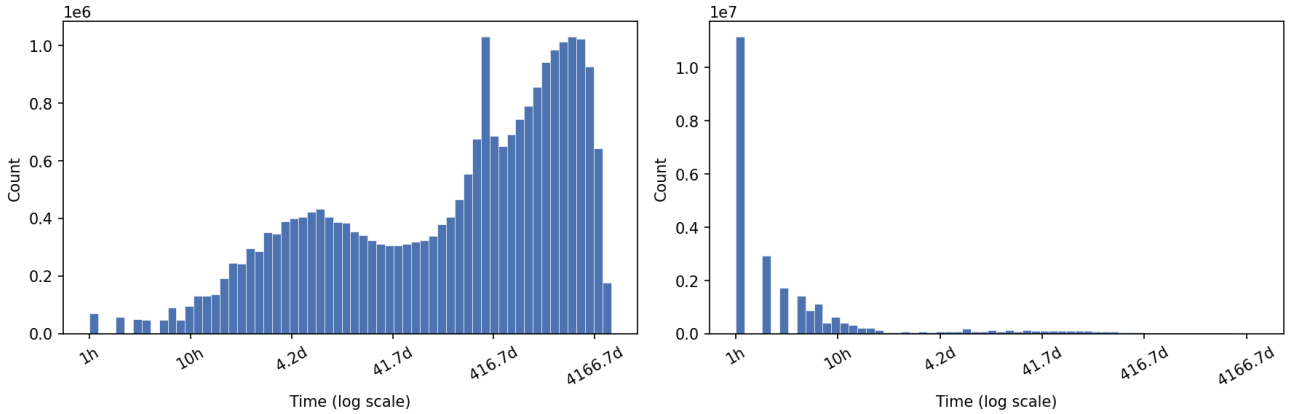

Figure 3. **Distribution of temporal characteristics across the MIMIC-IV population.** (Left) Histogram of elapsed time in hours from each medical event to the start of the patient sequence. (Right) Histogram of inter-event gaps plotted on a log scale. Zero-duration intervals are excluded.

A.3. Time Token Binning Strategy

Table 2. Discrete time token bin boundaries. Elapsed time in hours is mapped to one of ten tokens based on the bin in which it falls.

Token	Elapsed time range (hours)
1	2 – 5
2	6 – 11
3	12 – 23
4	24 – 95
5	96 – 215
6	216 – 455
7	456 – 839
8	840 – 1,678
9	1,680 – 3,599
10	3,600 – 103,794

A.4. Downstream Performance When Truncating Medical History

This ablation study explores the relative performance of the different models when historical data from previous visits is removed. Therefore, the model is only fed the first 48 hours of the current hospital admission.

Similarly to the results with historical visits included, the inclusion of time does not show a dramatic change in model performance. This is to be expected, as time information is likely most informative to the model when deciding the relevance

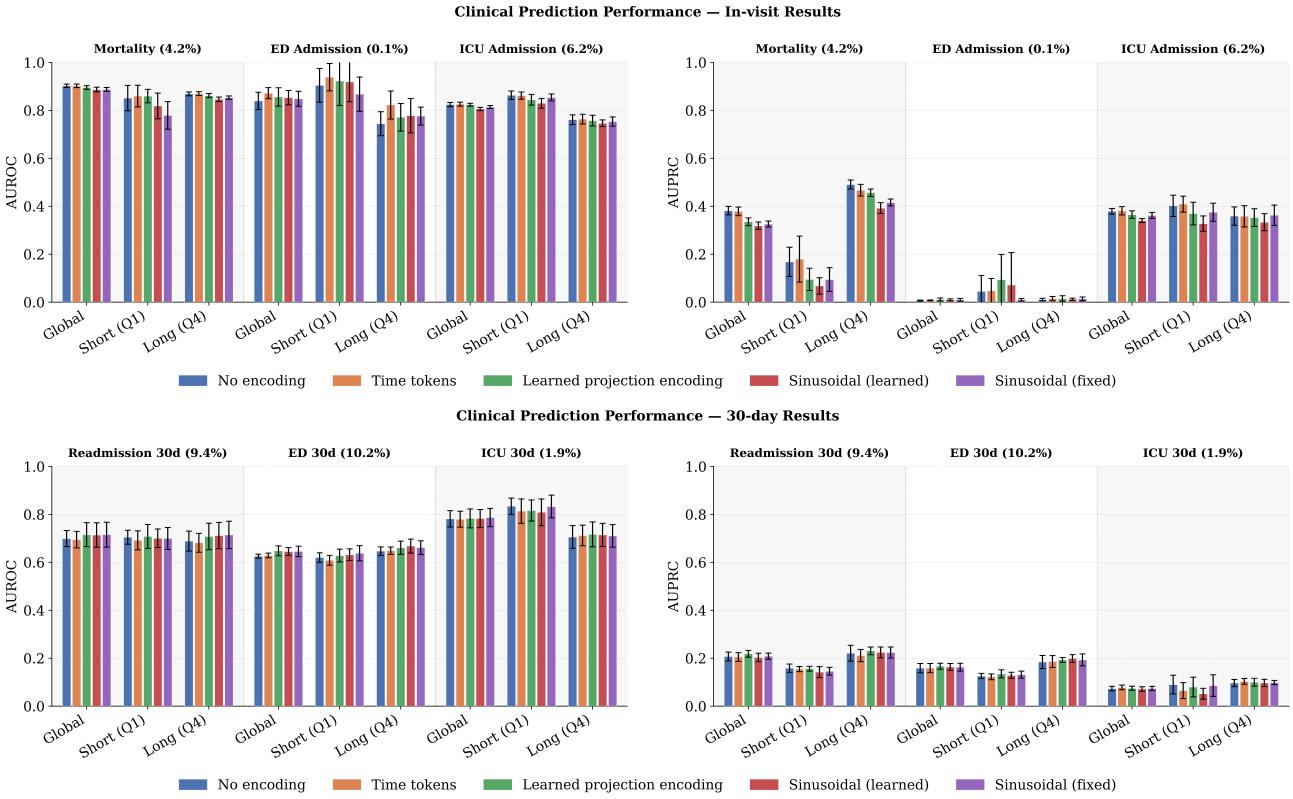

Figure 4. **Downstream Performance of GRU foundation model final embeddings under different time encoding conditions using only 48 hours of the current visit's data.** Comparison of AUROC and AUPRC (mean across 10 seeds ± 95% confidence interval) across six clinical prediction tasks. All historical events preceding the randomly chosen evaluation visit are removed to emulate the first presentation of the patient to the healthcare service.

of historical events, where distant events may be down-weighted, indicating the recovery from a medical condition. To explore this hypothesis a further ablation was considered where the model only receives historical medical data from previous visits, and attempts to predict whether events occur in, or following, the current visit.

A.5. Downstream Performance Only Including Medical History

This ablation study explores the relative performance of the different models when historical data from previous visits is included, but all information from the current visit is removed. This results in a larger cohort of eligible individuals, as there is no truncation due to potential data leakage, producing a much larger prevalence for the in-visit ED and ICU admission tasks.

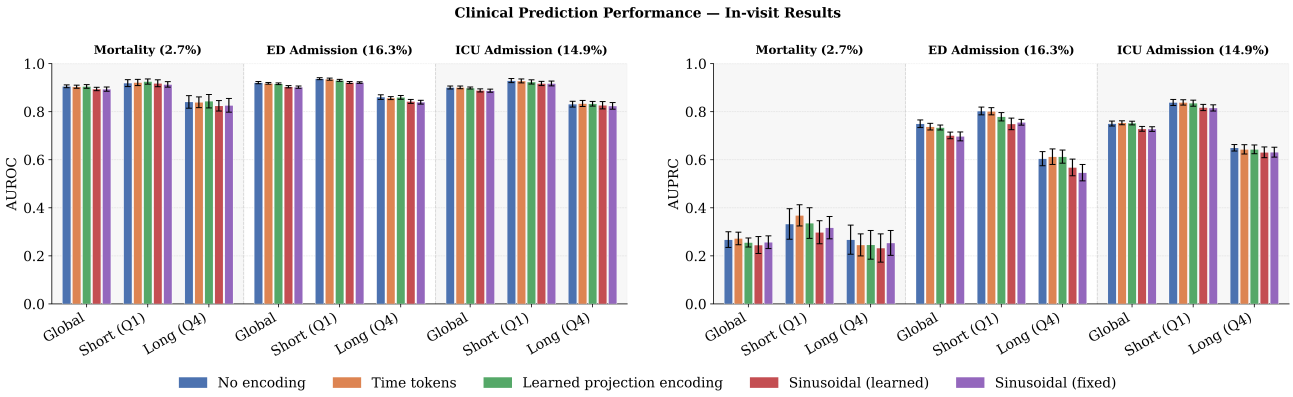

Figure 5. **Downstream Performance of GRU foundation model final embeddings under different time encoding conditions with only prior-visit data.** Comparison of AUROC and AUPRC (mean across 10 seeds ± 95% confidence interval) across six clinical prediction tasks. All in-visit events within the randomly chosen evaluation visit are removed to emulate the presentation of the patient to the healthcare service with only their historical data available.

A.6. Raw Results

Table 3. Downstream clinical prediction performance (AUROC / AUPRC, mean ± 95% CI across 10 seeds) for all time encoding conditions, stratified by sequence length, using full patient history. Prevalence of the positive class in the global stratum is shown in parentheses.

		Global				Short (Q1)				Long (Q4)			
Task	**Condition**	AUROC	±	AUPRC	±	AUROC	±	AUPRC	±	AUROC	±	AUPRC	±
Mortality (4.0%)	No encoding	**0.918**	0.006	**0.429**	0.023	0.902	0.030	0.172	0.052	**0.890**	0.010	**0.437**	0.041
	Time tokens	0.917	0.006	0.379	0.035	**0.907**	0.032	0.183	0.058	0.884	0.008	0.388	0.054
	Learned	0.915	0.006	0.362	0.036	0.907	0.030	**0.189**	0.063	0.879	0.011	0.375	0.051
	Sinus. (fixed)	0.910	0.005	0.345	0.025	0.892	0.033	0.130	0.048	0.873	0.009	0.357	0.047
	Sinus. (learned)	0.903	0.006	0.336	0.030	0.862	0.032	0.142	0.070	0.867	0.008	0.358	0.046
ED in visit (0.4%)	No encoding	0.951	0.016	**0.701**	0.068	0.992	0.008	**0.915**	0.047	0.930	0.044	**0.616**	0.204
	Time tokens	0.938	0.022	0.691	0.069	0.985	0.011	0.890	0.056	0.913	0.058	0.586	0.219
	Learned	0.952	0.015	0.662	0.083	**0.995**	0.004	0.894	0.082	0.919	0.057	0.586	0.219
	Sinus. (fixed)	0.948	0.012	0.682	0.067	0.988	0.011	0.869	0.087	0.921	0.046	0.609	0.205
	Sinus. (learned)	**0.953**	0.017	0.680	0.054	0.985	0.011	0.879	0.061	**0.940**	0.035	0.600	0.217
ICU in visit (5.8%)	No encoding	0.867	0.008	0.501	0.029	0.947	0.007	0.733	0.039	0.770	0.026	**0.222**	0.043
	Time tokens	**0.869**	0.008	**0.504**	0.027	**0.957**	0.006	**0.766**	0.033	0.753	0.021	0.187	0.039
	Learned	0.867	0.008	0.489	0.028	0.946	0.011	0.737	0.038	**0.777**	0.024	0.198	0.038
	Sinus. (fixed)	0.860	0.009	0.485	0.030	0.941	0.011	0.718	0.043	0.756	0.020	0.185	0.022
	Sinus. (learned)	0.854	0.007	0.483	0.034	0.937	0.012	0.718	0.045	0.751	0.024	0.200	0.042
Readmission 30d (9.3%)	No encoding	**0.760**	0.059	0.255	0.023	**0.744**	0.047	0.162	0.066	0.686	0.026	0.319	0.049
	Time tokens	0.751	0.060	0.258	0.031	0.717	0.032	0.160	0.067	0.681	0.025	0.324	0.046
	Learned	0.759	0.062	0.251	0.030	0.741	0.066	0.145	0.054	0.694	0.028	0.319	0.044
	Sinus. (fixed)	0.760	0.065	**0.261**	0.033	0.729	0.059	0.148	0.054	**0.697**	0.032	**0.330**	0.052
	Sinus. (learned)	0.757	0.064	0.259	0.030	0.736	0.065	**0.164**	0.071	0.693	0.029	0.323	0.044
ED 30d (10.1%)	No encoding	0.672	0.025	0.186	0.014	0.630	0.016	0.107	0.023	0.627	0.029	0.237	0.028
	Time tokens	0.674	0.025	0.189	0.013	0.625	0.013	0.103	0.020	0.625	0.026	0.236	0.027
	Learned	**0.678**	0.031	**0.193**	0.015	0.633	0.012	0.112	0.024	0.634	0.036	0.245	0.031
	Sinus. (fixed)	0.678	0.029	0.193	0.017	**0.638**	0.018	**0.113**	0.025	**0.637**	0.034	**0.249**	0.033
	Sinus. (learned)	0.675	0.030	0.187	0.013	0.637	0.016	0.110	0.021	0.626	0.037	0.232	0.029
ICU 30d (1.8%)	No encoding	**0.816**	0.037	0.093	0.016	0.823	0.054	0.073	0.057	0.721	0.027	0.117	0.023
	Time tokens	0.816	0.036	0.091	0.016	**0.826**	0.078	0.071	0.050	**0.726**	0.021	0.121	0.029
	Learned	0.811	0.033	**0.098**	0.019	0.825	0.057	0.068	0.054	0.716	0.020	**0.128**	0.031
	Sinus. (fixed)	0.815	0.040	0.090	0.019	0.817	0.089	**0.096**	0.057	0.717	0.025	0.117	0.026
	Sinus. (learned)	0.809	0.037	0.090	0.014	0.804	0.080	0.077	0.052	0.715	0.020	0.121	0.025

Table 4. Downstream clinical prediction performance (AUROC / AUPRC, mean ± 95% CI across 10 seeds) for all time encoding conditions, stratified by sequence length, using only admission data (no prior visit history). Prevalence of the positive class in the global stratum is shown in parentheses.

		Global				Short (Q1)				Long (Q4)			
Task	**Condition**	AUROC	±	AUPRC	±	AUROC	±	AUPRC	±	AUROC	±	AUPRC	±
Mortality (4.2%)	No encoding	**0.903**	0.007	**0.382**	0.018	0.852	0.053	0.168	0.061	0.869	0.008	**0.491**	0.019
	Time tokens	0.902	0.007	0.379	0.018	**0.860**	0.046	**0.180**	0.096	**0.870**	0.008	0.467	0.024
	Learned	0.896	0.008	0.335	0.016	0.860	0.028	0.095	0.047	0.862	0.009	0.457	0.016
	Sinus. (fixed)	0.887	0.008	0.326	0.013	0.779	0.058	0.094	0.050	0.853	0.007	0.416	0.014
	Sinus. (learned)	0.887	0.010	0.319	0.016	0.818	0.054	0.068	0.035	0.846	0.009	0.392	0.023
ED in visit (0.1%)	No encoding	0.840	0.036	0.007	0.002	0.905	0.070	0.045	0.067	0.744	0.050	0.011	0.005
	Time tokens	**0.872**	0.023	0.008	0.002	**0.939**	0.057	0.047	0.052	**0.823**	0.058	**0.015**	0.009
	Learned	0.856	0.038	**0.011**	0.006	0.923	0.102	**0.094**	0.105	0.771	0.057	0.015	0.012
	Sinus. (fixed)	0.849	0.031	0.010	0.006	0.868	0.071	0.010	0.006	0.776	0.038	0.014	0.008
	Sinus. (learned)	0.853	0.030	0.010	0.004	0.919	0.083	0.072	0.135	0.778	0.072	0.012	0.005
ICU in visit (6.2%)	No encoding	0.824	0.009	0.379	0.012	**0.863**	0.018	0.402	0.045	0.761	0.021	0.359	0.038
	Time tokens	**0.826**	0.008	**0.381**	0.017	0.862	0.015	**0.409**	0.033	**0.764**	0.020	0.358	0.044
	Learned	0.824	0.006	0.365	0.016	0.844	0.022	0.370	0.047	0.757	0.022	0.353	0.037
	Sinus. (fixed)	0.814	0.006	0.362	0.013	0.854	0.014	0.375	0.038	0.754	0.019	**0.363**	0.043
	Sinus. (learned)	0.806	0.007	0.340	0.008	0.830	0.021	0.328	0.032	0.746	0.014	0.334	0.036
Readmission 30d (9.4%)	No encoding	0.699	0.033	0.208	0.018	0.705	0.029	**0.158**	0.017	0.689	0.042	0.221	0.033
	Time tokens	0.695	0.034	0.205	0.018	0.692	0.040	0.155	0.011	0.682	0.040	0.211	0.026
	Learned	**0.716**	0.050	**0.218**	0.014	**0.708**	0.050	0.156	0.010	0.708	0.055	**0.231**	0.016
	Sinus. (fixed)	0.716	0.052	0.209	0.013	0.700	0.046	0.146	0.016	**0.715**	0.057	0.224	0.023
	Sinus. (learned)	0.714	0.051	0.203	0.018	0.700	0.039	0.142	0.023	0.712	0.055	0.224	0.023
ED 30d (10.2%)	No encoding	0.626	0.008	0.159	0.020	0.620	0.020	0.126	0.011	0.647	0.018	0.184	0.027
	Time tokens	0.629	0.010	0.159	0.019	0.609	0.021	0.122	0.012	0.649	0.015	0.186	0.025
	Learned	**0.648**	0.021	**0.167**	0.013	0.628	0.027	**0.135**	0.017	0.661	0.027	0.193	0.010
	Sinus. (fixed)	0.645	0.022	0.163	0.016	**0.638**	0.032	0.131	0.015	0.661	0.028	0.193	0.025
	Sinus. (learned)	0.645	0.017	0.164	0.014	0.632	0.025	0.129	0.013	**0.668**	0.029	**0.199**	0.016
ICU 30d (1.9%)	No encoding	0.781	0.034	0.073	0.010	**0.834**	0.034	**0.090**	0.039	0.706	0.047	0.097	0.015
	Time tokens	0.780	0.033	**0.078**	0.010	0.814	0.051	0.065	0.033	0.712	0.043	**0.104**	0.012
	Learned	0.783	0.040	0.075	0.009	0.816	0.044	0.080	0.041	**0.717**	0.052	0.100	0.017
	Sinus. (fixed)	**0.787**	0.038	0.074	0.009	0.833	0.048	0.086	0.045	0.711	0.047	0.099	0.009
	Sinus. (learned)	0.783	0.038	0.072	0.010	0.809	0.055	0.052	0.023	0.715	0.048	0.097	0.016

Table 5. Downstream clinical prediction performance (AUROC / AUPRC, mean ± 95% CI across 10 seeds) for all time encoding conditions, stratified by sequence length, using only prior visit history. Prevalence of the positive class in the global stratum is shown in parentheses.

		Global				Short (Q1)				Long (Q4)			
Task	**Condition**	AUROC	±	AUPRC	±	AUROC	±	AUPRC	±	AUROC	±	AUPRC	±
Mortality (2.7%)	No encoding	**0.9056**	0.0061	0.2675	0.0327	0.9190	0.0145	0.3326	0.0635	0.8409	0.0262	**0.2674**	0.0604
	Time tokens	0.9040	0.0066	**0.2725**	0.0264	0.9217	0.0130	**0.3686**	0.0442	0.8394	0.0223	0.2457	0.0457
	Learned	0.9052	0.0076	0.2556	0.0188	**0.9254**	0.0108	0.3364	0.0640	**0.8437**	0.0278	0.2461	0.0598
	Sinus. (fixed)	0.8938	0.0089	0.2569	0.0264	0.9136	0.0115	0.3176	0.0464	0.8265	0.0284	0.2539	0.0517
	Sinus. (learned)	0.8944	0.0069	0.2452	0.0353	0.9183	0.0145	0.2983	0.0481	0.8246	0.0217	0.2326	0.0587
ED in visit (16.3%)	No encoding	**0.9210**	0.0047	**0.7496**	0.0160	**0.9373**	0.0042	**0.8030**	0.0161	**0.8608**	0.0097	0.6045	0.0294
	Time tokens	0.9180	0.0037	0.7376	0.0137	0.9354	0.0045	0.8018	0.0148	0.8567	0.0061	0.6127	0.0324
	Learned	0.9164	0.0034	0.7341	0.0105	0.9303	0.0039	0.7790	0.0173	0.8597	0.0079	**0.6129**	0.0272
	Sinus. (fixed)	0.9022	0.0048	0.6972	0.0185	0.9213	0.0032	0.7559	0.0120	0.8397	0.0080	0.5462	0.0342
	Sinus. (learned)	0.9041	0.0046	0.7012	0.0137	0.9213	0.0041	0.7490	0.0241	0.8425	0.0082	0.5680	0.0349
ICU in visit (14.9%)	No encoding	0.9008	0.0062	0.7506	0.0101	**0.9294**	0.0087	0.8384	0.0124	0.8314	0.0121	**0.6495**	0.0135
	Time tokens	**0.9010**	0.0055	**0.7539**	0.0084	0.9280	0.0080	**0.8386**	0.0112	**0.8342**	0.0129	0.6432	0.0195
	Learned	0.8987	0.0042	0.7527	0.0074	0.9238	0.0089	0.8354	0.0129	0.8332	0.0097	0.6433	0.0181
	Sinus. (fixed)	0.8872	0.0064	0.7278	0.0097	0.9175	0.0099	0.8157	0.0130	0.8240	0.0138	0.6313	0.0205
	Sinus. (learned)	0.8886	0.0064	0.7286	0.0100	0.9173	0.0089	0.8175	0.0131	0.8267	0.0154	0.6310	0.0223