Pretraining EHR Foundation Models with Patient-Aware Sampling

Anonymous Authors[1]

Abstract

Autoregressive foundation models for electronic health records (EHRs) typically inherit pretraining methods from language modeling, where patient trajectories are concatenated into a global token stream and windows are sampled from that stream. In EHR data, this choice is consequential: windows may mix multiple patients, and patients with longer records contribute more optimization updates, potentially biasing learning toward long trajectories. We propose alternative pretraining sequence construction methods, focusing on how training signal is distributed across patients. Specifically, we compare Global Stream, deterministic Patient Chunks, and stochastic Patient Sampling with controllable weighting. Across downstream clinical tasks on MIMIC-IV v2.2 and v3.1, Patient Sampling improves Macro AUROC and AUPRC over the Global Stream baseline. These results identify training and validation sequence construction as important and underexplored design choices for autoregressive EHR foundation models.

1. Introduction

Autoregressive foundation models for electronic health records (EHRs) represent each patient history as a variable-length sequence of discrete tokens and train with a next-token prediction objective. Recent work has shown that these models can support zero-shot clinical prediction by simulating future trajectories from observed patient histories (Renc et al., 2024; Kraljevic et al., 2024).

Current EHR pretraining methods largely inherit sequence construction strategies from language modeling. In the standard setup, patient trajectories are concatenated into a global token stream and fixed-length training windows are sampled from that stream. In EHR data, this is more than an efficiency choice: a window may contain events from multiple patients, and patients with longer records contribute more optimization updates during training. This is especially relevant in our setting, where patient trajectory lengths are highly unequal, forming a log-normal distribution as shown in Figure 3.

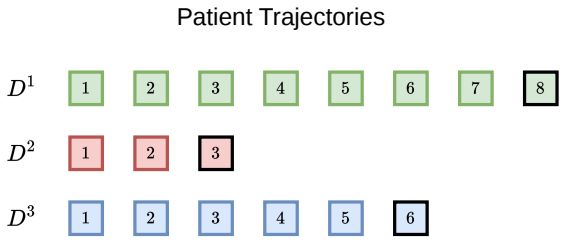

Figure 1. Toy example of an EHR dataset represented as variable-length token sequences D_i. Colored boxes denote EHR tokens, and black-bordered boxes indicate end-of-sequence tokens.

In this work, we study alternative methods for pretraining sequence construction. Keeping tokenization and model architecture fixed, we compare the standard Global Stream method with patient-aware alternatives that preserve patient boundaries and alter how training signal is distributed across patients. Across MIMIC-IV v2.2 and v3.1 (Johnson et al., 2023b; 2024; 2023a), we find that Patient Sampling achieves stronger overall downstream performance than the Global Stream baseline, improving both Macro AUROC and Macro AUPRC. These findings highlight pretraining sequence construction as an important and previously underexplored design choice in EHR foundation models.

2. Background

Autoregressive modeling has recently emerged as a promising paradigm for EHR data, where patient history is represented as a sequence of tokenized events. Figure 1 illustrates this setup with a toy tokenized EHR dataset.

EHR Foundation Models Recent work has applied autoregressive or generative pretraining to longitudinal EHR data. ETHOS and its follow-up ETHOS-ARES convert EHR data into tokenized patient timelines, train GPT-2 (Radford et al., 2019) style models with next-token prediction, and

[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.

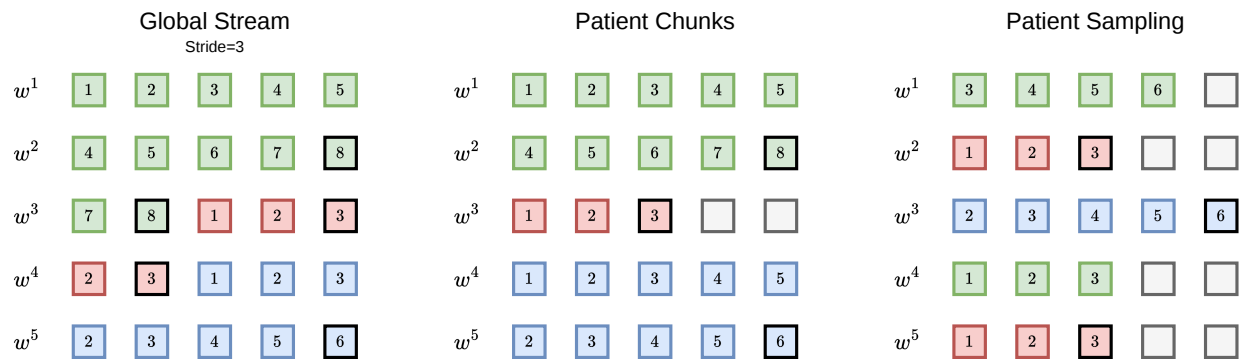

Figure 2. Comparison of the dataset constructions considered in this work. On the left *Global Stream* forms training windows from a global concatenated token stream. *Patient Chunks* constructs deterministic within-patient windows. *Patient Sampling* first samples a patient and then samples a window from that patient alone. Colors denote different patients, and gray boxes indicate padding.

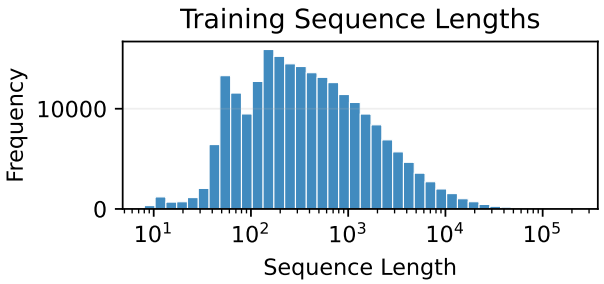

Figure 3. Distribution of patient token sequence lengths in the MIMIC-IV v2.2 training set, shown on a log-scaled x-axis.

evaluate them by rolling out future events from a clinical prediction point to obtain zero-shot predictions for downstream tasks (Renc et al., 2024; 2025). Foresight models patient timelines with a generative pretrained transformer that combines structured EHR data and free text to predict future clinical concepts and outcomes (Kraljevic et al., 2024). CoMET shows that autoregressive medical event pretraining with a next-token prediction objective scales to larger models and datasets (Waxler et al., 2025). Together, these works establish autoregressive pretraining over patient timelines as a viable approach to EHR foundation modeling, but do not explicitly study pretraining sequence construction as a primary design variable.

Sequence Composition Recent work in language model pretraining suggests that sequence composition is an important design choice. Zhao et al. (2024) show that concatenating unrelated documents into a fixed-length sequence can introduce distracting cross-document context, harming language modeling and downstream performance. They demonstrate that both intra-document masking and retrieval-based packing of related documents improve performance. Similarly, In-Context Pretraining shows that re-ordering documents so that each context window contains semantically related material improves tasks requiring stronger contextual reasoning across document boundaries (Shi et al., 2024). These findings motivate studying sequence composition in EHR modeling, where patient trajectories are highly unequal in length and standard sequence construction may further amplify this imbalance.

3. Method

We consider a dataset of N patients $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$, where patient i is represented by a sequence of EHR tokens

$$x^{(i)} = \left(x_1^{(i)}, \ldots, x_{L_i}^{(i)}\right),$$

and L_i is the sequence length. Following prior work on autoregressive EHR modeling, we train a decoder-only transformer with a next-token prediction objective on fixed-length token windows of length S.

We compare several strategies for constructing windows $w = (w_1, \ldots, w_S)$ for training and validation.

Global Stream. Our baseline follows the standard global stream construction used in autoregressive pretraining. We concatenate all patient sequences into a single stream

$$z = x^{(1)} \parallel x^{(2)} \parallel \cdots \parallel x^{(N)},$$

with total length

$$T = \sum_{i=1}^{N} L_i.$$

Given stride r, we define the set of valid start indices

$$\mathcal{T}_{\mathrm{GS}} = \{1,\ 1{+}r,\ 1{+}2r,\ \ldots,\ t_{\max}\}, \qquad t_{\max} \le T{-}S{+}1.$$

Each training example is then

$$w^{(t)} = (z_t, \ldots, z_{t+S-1}), \qquad t \in \mathcal{T}_{\mathrm{GS}}.$$

This construction is simple and efficient, but a single window may contain tokens from multiple patients.

Dataset Method			MIMIC-IV v2.2 + ED v2.2				MIMIC-IV v3.1 + ED v2.2					
Train	Validation	α	ICU Mortality		ICU Readmission		ICU Mortality		ICU Readmission		Macro	
			AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Global Stream	Global Stream	–	0.8781	0.4338	0.725	0.383	0.843	0.428	0.703	0.380	0.787	0.406
Global Stream	Patient Chunks	–	0.866	0.428	0.717	0.389	**0.852**	0.439	0.712	0.382	0.786	0.410
Patient Chunks	Patient Chunks	–	0.865	0.357	0.664	0.313	0.832	0.327	0.662	0.334	0.756	0.332
Patient Sampling	Patient Chunks	0	**0.885**	0.432	0.726	0.392	0.851	0.437	0.701	0.372	0.791	0.408
		0.2	0.848	0.350	0.721	0.382	0.839	0.402	0.690	0.365	0.774	0.375
		0.4	0.882	0.443	0.728	0.399	0.834	0.405	0.696	0.372	0.785	0.405
		0.6	0.884	**0.456**	0.730	0.405	0.850	**0.446**	0.706	0.381	**0.792**	**0.422**
		0.8	0.845	0.378	0.725	0.397	0.821	0.379	0.707	0.376	0.775	0.383
		1	0.863	0.416	**0.733**	**0.407**	0.848	0.435	**0.719**	**0.399**	0.791	0.414

Table 1. Comparison of pretraining data constructions on ICU Mortality and ICU Readmission across MIMIC-IV v2.2 + ED v2.2 and MIMIC-IV v3.1 + ED v2.2. Results are reported as AUROC and AUPRC. The final macro column denotes the mean across all four reported task settings, and the best value in each metric column is shown in bold.

Patient Chunks. To preserve patient boundaries, we construct windows independently within each patient sequence. For patient i, we define a deterministic set of chunk start indices

$$\mathcal{B}_i = \{1,\ 1+S,\ 1+2S,\ \ldots,\ b_{i,\max}\},$$

where the final chunk is right-aligned to the end of the sequence,

$$b_{i,\max} = \max(1,\ L_i - S + 1).$$

Each chunk is

$$w^{(i,b)} = \left(x_b^{(i)}, \ldots, x_{\min(L_i,\, b+S-1)}^{(i)}\right), \qquad b \in \mathcal{B}_i.$$

Chunks shorter than S are right-padded to length S, and padded target positions are excluded from the loss. Unlike Global Stream, Patient Chunks always produces windows containing tokens from a single patient.

Patient Sampling. We next introduce a stochastic construction that decouples *patient* and *window* selection. This construction is only used for training due to its non-deterministic nature.

For each patient i, let $\mathcal{W}^{(i)}$ denote the set of valid start indices for constructing a window. We first sample a patient according to

$$p_\alpha(i) = \frac{|\mathcal{W}^{(i)}|^\alpha}{\sum_{j=1}^{N} |\mathcal{W}^{(j)}|^\alpha},$$

where $\alpha \in [0, 1]$ controls the weighting of patients during training. When $\alpha = 0$, patients are sampled uniformly; when $\alpha = 1$, patients are sampled proportionally to the number of valid windows they contain, i.e. a patient with twice as many possible token windows as another patient will be sampled twice as often.

Given a sampled patient i, we sample a start index uniformly from

$$\mathcal{W}^{(i)} = \{1 - S,\ 2 - S,\ \ldots,\ L_i - 1\}.$$

This choice of start indices allows tokens within a patient trajectory to appear with a broader range of left-context lengths, rather than systematically under-sampling early positions.[1] Conditioned on (i, s), the resulting window is $x_{s:s+S-1}^{(i)}$, right-padded to length S if necessary. Patient Sampling preserves patient boundaries while allowing the contribution of individual patients to be controlled continuously through α.

Global Stream may cross patient boundaries and over-represent long trajectories, Patient Chunks preserves patient boundaries but retains this imbalance, and Patient Sampling preserves boundaries while allowing the training distribution over patients to vary smoothly between patient-uniform and length-weighted sampling. Each method is visualized in Figure 2.

4. Experiments and Results

Experimental setup. We evaluate the methods introduced in Section 3 using two tokenized EHR datasets derived from MIMIC-IV: version 2.2 (Johnson et al., 2023b) and version 3.1 (Johnson et al., 2024), both accessed via PhysioNet (Goldberger et al., 2000). In both settings, we also include the Emergency Department module, MIMIC-IV-ED version 2.2 (Johnson et al., 2023a). Each dataset is split by patient into 80% train, 10% validation, and 10% test sets.

We train a 6-layer GPT-2 model (Radford et al., 2019) on a single H100 with sequence length $S = 2048$. Validation loss is evaluated every 10K training steps, and the checkpoint with the lowest validation loss is used for downstream evaluation. Because Patient Chunks is deterministic and

[1] Due to the need for an input and a target, the first and last tokens are still slightly under-sampled compared to interior tokens, though we believe this effect is minimal.

Method	ICU Mortality		ICU Readmission		ICU Admission		Hospital Mortality		Macro	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
MIMIC-IV v2.2 + ED v2.2										
Global Stream	0.878	0.434	0.725	0.383	**0.909**	**0.760**	0.826	0.289	0.834	0.467
Patient Sampling	**0.884**	**0.456**	**0.730**	**0.405**	**0.909**	0.756	**0.846**	**0.297**	**0.842**	**0.478**
MIMIC-IV v3.1 + ED v2.2										
Global Stream	0.843	0.428	0.703	0.380	0.905	0.748	0.813	0.308	0.816	0.466
Patient Sampling	**0.850**	**0.446**	**0.706**	**0.381**	**0.910**	**0.757**	**0.834**	**0.322**	**0.825**	**0.477**

Table 2. Comparison of Global Stream and Patient Sampling ($\alpha = 0.6$) across ICU Mortality, ICU Readmission, ICU Admission, and Hospital Mortality for MIMIC-IV v2.2 + ED v2.2 and MIMIC-IV v3.1 + ED v2.2. Results are reported as AUROC and AUPRC, and the Macro columns report the mean across the four tasks within each dataset version. Best value in each column is shown in bold.

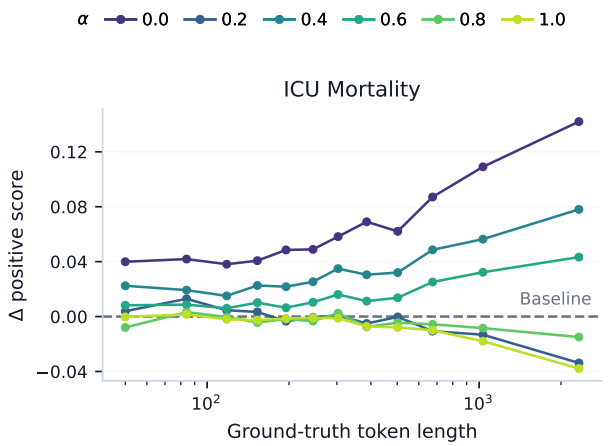

Figure 4. Δ positive score on ICU Mortality (MIMIC-IV v2.2) for Patient Sampling models with different α values, relative to a Global Stream baseline. Positive score is the fraction of simulations that terminate with the correct end token, delta positive score is measured relative to the Global Stream baseline. Datapoints are grouped by the number of tokens of the ground truth sequence.

both patient-aware methods operate on single-patient windows, we use Patient Chunks validation for Patient Chunks and Patient Sampling. For Global Stream, we report results with both Global Stream and Patient Chunks validation. We evaluate downstream performance using the rollout framework of Renc et al. (2025) on the following clinical tasks: ICU Mortality, ICU Readmission, ICU Admission, and Hospital Mortality. See Appendix A for full training and evaluation details.

Main results. We first evaluate each method on the ICU Mortality and ICU Readmission tasks, sweeping Patient Sampling over α values in increments of 0.2. Results are shown in Table 1. We find that Patient Chunks does not improve over the Global Stream baseline, suggesting that simply preserving patient boundaries is not sufficient. Patient Sampling with $\alpha = 0.6$ yields the strongest macro performance across both tasks in both datasets, and we therefore select this model for broader evaluation.

Table 2 shows results from further evaluation on the Hospital Mortality and ICU Admission tasks. We find performance improvements on the Hospital Mortality task across both datasets. Improvements on ICU Admission are marginal, with no gain on MIMIC-IV v2.2 and small gains on MIMIC-IV v3.1. Across all four benchmarks and both datasets, Patient Sampling outperforms Global Stream on all four macro metrics.

Effect of α. Figure 4 shows how the behavior of Patient Sampling varies with ground-truth token length on the ICU Mortality benchmark for MIMIC-IV v2.2. Smaller values of α produce the strongest positive shifts relative to the Global Stream baseline, with $\alpha = 0$ showing the largest effect overall. As hypothesized, $\alpha = 0$ improves performance on shorter sequences, but this effect is actually amplified as ground-truth sequence length increases. Intermediate settings such as $\alpha = 0.4$ and $\alpha = 0.6$ also remain positive across much of the range, whereas larger values of α stay closer to parity or become negative as trajectory length increases. Notably, $\alpha = 0.2$ underperforms relative to neighboring values of α, highlighting the need for further investigation into the effect of α and the behavior of the learned model.

5. Conclusion

We studied how pretraining sequence construction affects autoregressive EHR foundation models, proposing a new Patient Sampling method. We show Patient Sampling improves on the Global Stream baseline across a broad set of downstream benchmarks. These results suggest that patient boundaries alone are insufficient, but that patient-aware sampling and control over how training signal is distributed can improve autoregressive EHR pretraining. More broadly, our findings identify sequence construction as an important design choice for autoregressive EHR foundation models.

References

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. RRID:SCR_007345.

Johnson, A., Bulgarelli, L., Pollard, T., Celi, L. A., Mark, R., and Horng, S. Mimic-iv-ed (version 2.2), 2023a. URL https://doi.org/10.13026/5ntk-km72. RRID:SCR_007345.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark, R. MIMIC-IV. *PhysioNet*, January 2023b. doi: 10.13026/6mm1-ek67. URL https://doi.org/10.13026/6mm1-ek67. Version 2.2.

Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., and Mark, R. Mimic-iv (version 3.1), 2024. URL https://doi.org/10.13026/kpb9-mt58. RRID:SCR_007345.

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, 2024. doi: 10.1016/S2589-7500(24)00025-6. URL https://doi.org/10.1016/S2589-7500(24)00025-6.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

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

Renc, P., Grzeszczyk, M. K., Oufattole, N., Goode, D., Jia, Y., Bieganski, S., McDermott, M. B. A., Was, J., Samir, A. E., Cunningham, J. W., Bates, D. W., and Sitek, A. Foundation model of electronic medical records for adaptive risk estimation. *GigaScience*, 14:giaf107, 09 2025. ISSN 2047-217X. doi: 10.1093/gigascience/giaf107. URL https://doi.org/10.1093/gigascience/giaf107.

Shi, W., Min, S., Lomeli, M., Zhou, C., Li, M., Lin, X. V., Smith, N. A., Zettlemoyer, L., tau Yih, W., and Lewis, M. In-context pretraining: Language modeling beyond document boundaries. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LXVswInHOo.

Waxler, S., Blazek, P., White, D., Sneider, D., Chung, K., Nagarathnam, M., Williams, P., Voeller, H., Wong, K., Swanhorst, M., Zhang, S., Usuyama, N., Wong, C., Naumann, T., Poon, H., Loza, A., Meeker, D., Hain, S., and Shah, R. Generative medical event models improve with scale, 2025. URL https://arxiv.org/abs/2508.12104.

Zhao, Y., Qu, Y., Staniszewski, K., Tworkowski, S., Liu, W., Miłoś, P., Wu, Y., and Minervini, P. Analysing the impact of sequence composition on language model pre-training. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7897–7912, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.427. URL https://aclanthology.org/2024.acl-long.427/.

Component	Setting
Embedding dimension	768
Transformer layers	6
Attention heads	12
Context length	2048
Dropout	0.1
RoPE base	10000
Batch size	64
Optimizer	AdamW
Learning rate	1.5×10^{-4}
Weight decay	0.1
Warmup steps	4000
Hold steps	136000
Decay steps	260000
Final LR ratio	0.05
Validation interval	every 10K steps

Table 3. Core model and training hyperparameters used in our experiments.

Benchmark	Prediction point	Terminating tokens
ICU Mortality	`ICU_ADMISSION`	`MEDS_DEATH`, `ICU_DISCHARGE`
ICU Readmission	`ICU_DISCHARGE`	`ICU_ADMISSION`, `HOSPITAL_DISCHARGE`
ICU Admission	`HOSPITAL_ADMISSION`	`ICU_ADMISSION`, `HOSPITAL_DISCHARGE`
Hospital Mortality	`HOSPITAL_ADMISSION`	`MEDS_DEATH`, `HOSPITAL_DISCHARGE`

Table 4. Benchmark definitions used in the rollout-based evaluation framework.

A. Model, Training, and Evaluation Framework

Model. We use a decoder-only transformer based on a GPT-2 style architecture. The model consists of a learned token embedding layer, a stack of transformer blocks with causal self-attention and MLP sublayers, a final layer normalization, and a linear language-modeling head. Rotary positional embeddings (RoPE) are used in place of learned absolute position embeddings. The model is trained with a standard next-token prediction objective over tokenized EHR trajectories. Table 3 summarizes the core model and optimization hyperparameters used in our experiments.

Evaluation framework. Downstream evaluation is performed using a rollout-based benchmark framework. For each benchmark, evaluation prompts are constructed by identifying a benchmark-specific prediction point together with two terminating outcome tokens in each patient trajectory. The model receives the patient history up to the prediction point and is then rolled out autoregressively until a terminating token is generated or a maximum generation length of 4096 tokens is reached. For each prompt, we run 20 stochastic rollouts. A scalar score is computed as the fraction of rollouts that terminate with the different end tokens, and this score is used as the model output for binary classification. Simulations that exceed the 4096-token limit are terminated and excluded from benchmark calculations. AUROC and AUPRC are then computed from the resulting per-prompt scores. Table 4 presents the definitions for each benchmark we report.