Analysis of heart failure patient trajectories using sequence modeling

Falk Dippel[a,*], Yinan Yu[b], Annika Rosengren[c,d], Martin Lindgren[c,d], Christina E. Lundberg[d,e], Erik Aerts[b], Martin Adiels[f], Helen Sjöland[c,d]

[a]*Sahlgrenska University Hospital, Diagnosvägen 11, Gothenburg, 416 85, Sweden*
[b]*Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Rännvägen 6B, Gothenburg, 412 96, Sweden*
[c]*Department of Molecular and Clinical Medicine, Sahlgrenska Academy, University of Gothenburg, Bruna stråket 16, Gothenburg, 413 45, Sweden*
[d]*Department of Medicine, Geriatrics and Emergency Medicine, Sahlgrenska University Hospital, Diagnosvägen 11, Gothenburg, 416 85, Sweden*
[e]*Department of Food and Nutrition, and Sport Science, Faculty of Education, University of Gothenburg, Box 300, Gothenburg, 405 30, Sweden*
[f]*School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Box 453, Gothenburg, 405 30, Sweden*

Abstract

Transformers have defined the state-of-the-art for clinical prediction tasks involving electronic health records (EHRs). The recently introduced Mamba architecture outperformed an advanced Transformer (Transformer++) based on Llama in handling long context lengths, while using fewer model parameters. Despite the impressive performance of these architectures, a systematic approach to empirically analyze model performance and efficiency under various settings is not well established in the medical domain. The performances of six sequence models were investigated across three architecture classes (Transformers, Transformers++, Mambas) in a large Swedish heart failure (HF) cohort (N = 42820), providing a clinically relevant case study. Patient data included diagnoses, vital signs, laboratories, medications and procedures extracted from in-hospital EHRs. The models were evaluated on three one-year prediction tasks: clinical instability (a readmission phenotype) after initial HF hospitalization, mortality after initial HF hospitalization and mortality after latest hospitalization. Ablations account for modifications of the EHR-based input patient sequence, architectural model configurations, and temporal preprocessing techniques for data collection. Llama achieves the highest predictive discrimination, best calibration, and showed robustness throughout all tasks across the conducted ablation studies, followed by Mambas. Both architectures demonstrate efficient representation learning, with tiny configurations surpassing other large-scaled Transformers. At equal model size, Llama and Mambas achieve superior performance using 25% less training data. This paper presents a first ablation study with systematic design choices for input tokenization, model configuration and temporal data preprocessing. Future model development in clinical prediction tasks using EHRs could build upon this study's recommendation as a starting point.

Keywords: Electronic health record; Transformer; Mamba; Heart failure; Mortality prediction

1. Introduction

Heart failure (HF) is a chronic, progressive condition characterized by impaired cardiac function, frequent hospitalizations, reduced quality of life and high mortality [1]. Prevalence increases steeply with age and HF often coexists and interacts causally with other cardiometabolic disorders and comorbidities [2, 3]. Personalized artificial intelligence (AI) systems have the potential for data-driven decision support in clinically relevant tasks guiding prediction-based interventions. If a patient with HF is predicted for mortality or clinical instability (a phenotype of readmission to hospital) within one-year with high confidence based on their available health information, patient-specific or outcome-specific interventions can be undertaken by the clinician or patient. For instance, a patient with a mortality prediction might be eligible for a tailored palliative care program prioritizing patient's quality of life and well-being over the strain caused by repeated hospitalizations. Alternatively, a patient with predicted clinical instability might be identified for early follow-up proactively anticipating and counteracting worsening or complications. Unreliable predictions place considerable demands as well as cost on the healthcare system and lead to suboptimal patient care decisions, motivating the development of highly precise AI models.

Patient health information is commonly stored and collected in longitudinal EHRs by healthcare providers, capturing a patient trajectory through irregular clinical visits, either hospitalizations or outpatient visits. An EHR includes structured data, e.g. recorded vital signs, results from laboratory tests and assessed diagnoses, as well as unstructured data, e.g. clinical text notes or medical images [4]. In recent years, the application of AI on EHRs has gained significant research interest due to the growing availability of EHRs and advances in AI [5, 6]. The field of natural language processing (NLP), in particular, has demonstrated significant success to learn patient representations from EHRs efficiently [6, 7, 8]. Transformers [9], the backbone architecture of most large language models (LLMs), have emerged in several healthcare applications, primarily clinical prediction tasks for structured EHRs [10], information extraction from clinical free-text notes [11] and medical imaging [12].

*Corresponding author: falk.dippel@vgregion.se

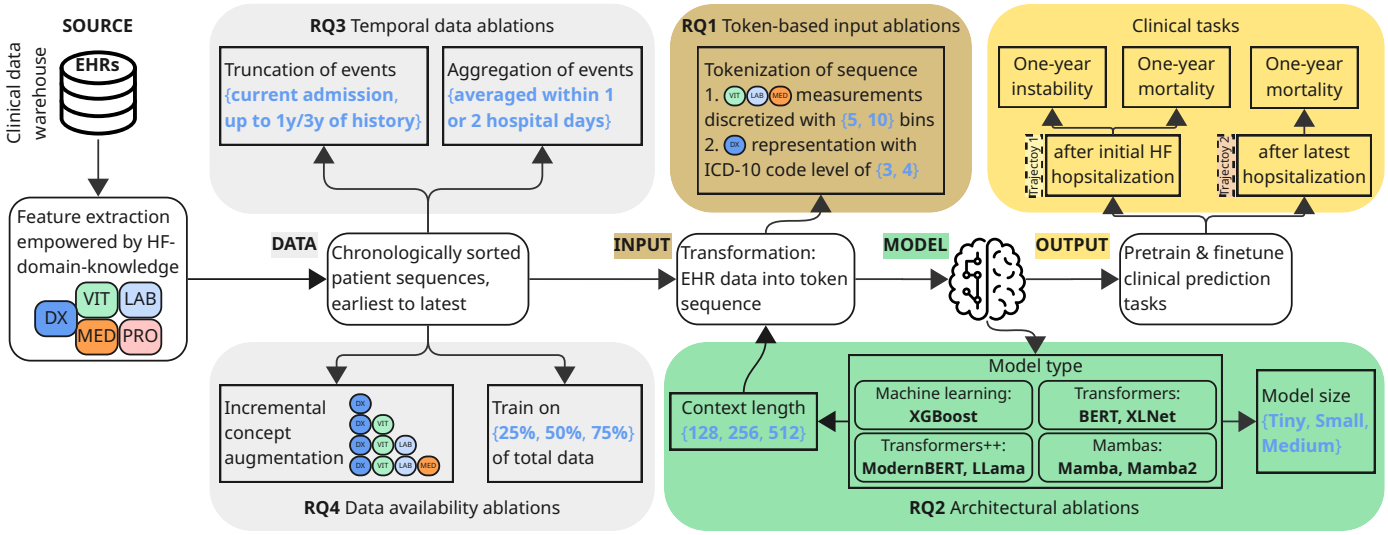


Fig. 1: Overview of the ablation study following a simplified development process, from source electronic health records (EHRs) through model training to clinical prediction: Research question **RQ1** explores variations of the tokenized input sequence, **RQ2** investigates the performance across different architectural configurations, **RQ3** analyzes temporal modifications of the data and **RQ4** investigates the model utility for varying data availability. Each ablation is conducted on three clinical tasks for all model types, with varying ablation parameters highlighted in blue. DX=Diagnoses. LAB=Laboratories. MED=Medications. PRO=Procedures. VIT=Vital signs.

Traditional deep learning (DL) approaches utilized convolutional neural networks [13] or recurrent neural networks (RNNs) with time attention mechanism [14] for EHR sequence modeling and prediction tasks. However, RNNs are known to struggle with long-range dependencies present in EHRs because of the vanishing gradient problem [15]. Although long short-term memory (LSTM) [16], an RNN variant with gated information flow, reduces the gradient problem, its complexity scales poorly, requiring significant computational resources or optimized training strategies [17]. Transformers, primarily BERT (bidirectional encoder representations from Transformers) [18], have demonstrated superior performance over RNNs [19, 20] and LSTMs [20, 21, 22] by capturing long-range dependencies and inter-concept relationships more efficiently through the attention mechanism [9]. Therefore, RNNs and LSTMs are not within the scope of this work.

Recent work has explored newer architectures (e.g., Llama [23] or Mamba [24]) which are capable of handling longer token sequences [25, 26]. The performance of Transformers scales with the dataset size (tokens) and model size (number of parameters) [27]. However, Transformers have a quadratic training time and memory complexity limiting their efficiency which arises from the self-attention mechanism learning pairwise contextual representations between all input tokens [9]. In contrast, Mamba is equipped with an attention-free design enabling linear scaling of input sequences due to the utilization of state space models (SSMs) empowered by a context-aware selection mechanism and hardware-aware algorithms which surpasses advanced Transformers (Transformers++) like Llama of larger scale [24, 28]. Transformers++ incorporate advanced architectural improvements (e.g., rotatory positional embeddings, pre-normalization, gating mechanisms) over Transformers [24].

Most previous works focus on general heterogeneous patient population [19, 29, 26]. In contrast, the present analysis focuses on EHRs from a large Swedish cohort with HF, a well-defined and validated condition [30]. We argue that disease-specific models tailored around defined diagnostic conditions are more likely to translate into clinical practice given that the availability and diversity of clinical concepts remain local or lack standardized EHRs. For HF outcomes, strong predictors are often concentrated in recent patient history, making shorter context lengths realistic, interpretable for clinicians and meaningful for sequence modeling. We tailor our study around shorter context lengths (≤ 512 tokens) and tiny- to medium-scaled model configurations (2 M to 30 M parameters) to deploy and to investigate sequence models for end users in a clinically realistic scenario with fewer computational resources. To the best of our knowledge, Mamba's predictive capability has not been thoroughly evaluated for shorter context lengths utilizing EHRs. Additionally, we explore two recently developed architectures that have not yet been introduced to EHR sequence modeling: ModernBERT [31], a BERT variant that has undergone several technical improvements similar to Llama, and Mamba2 [32], a more efficiently trainable Mamba variant. Lastly, we aim to provide insights into model behavior through a detailed ablation study, highlighting which design choices contribute to performance improvements. The components (Fig. 1) of our work summarizes as follows:

- Comparison of different sequence model architecture classes (Transformer, Transformer++, Mamba) across three one-year prediction tasks at different patient trajectories in a large Swedish HF cohort.

- Empirical analysis of input-centric, model-centric and data-centric ablations investigating the effects of varying token representations for clinical concepts, context length and architecture size, and temporal preprocessing techniques of the patient sequence.
- Data scalability analysis to examine performance changes with varying data availability of clinical concepts and training data.

Our main findings are: i) Llama outperforms both Mambas, which, in turn, significantly surpass the common BERT model at all studied context lengths (≤ 512 tokens) and model sizes, in most conducted ablations across all prediction tasks. ii) Llama and Mambas are efficient at representation learning as respective tiny-scaled variants outperform compared Transformers of larger scale. Moreover, Llama and Mamba require 25% less training data in our experiments to achieve convergence while outperforming BERT which was trained on all data.

2. Related work

Transformers, particularly BERT [18], have been widely applied to health data from EHRs for predictive tasks, as demonstrated by models such as BEHRT [19], BRLTM [20], and Med-BERT [29]. Previous works such as CEHR-BERT [21] and ExBEHRT [33] were pretrained on general patient populations and subsequently fine-tuned to predict 30-day readmission in HF subcohorts. Similarly, Med-BERT [29] was finetuned on patients with both diabetes and HF for disease prediction. Antikainen et al. [34] predicted 6-month mortality in cardiac patients using transformers. By incrementally optimizing BEHRT and applying techniques commonly found in Transformer++, CORE-BEHRT [35] identified substantial performance gains primarily from enhanced data representations incorporating medications and temporal encoding of patient age. CEHR-BERT [21] reformulated the temporal component of EHR representations by introducing an artificial time token [ATT] which stores the respective time gap between two consecutive visits, with each visit enclosed by a start [VS] and end token [VE]. To avoid the truncation of the input sequence by the context length, ExBEHRT [33] stacked clinical concepts vertically through additional embeddings. Similarly, Hi-BEHRT [36] introduced a slicing window mechanism to capture long-term dependencies between each visit segment. G-BERT [37] is based on a knowledge-driven approach where the ontology of medical concepts were learned through BERT embeddings combined with a graph neural network. Transformer-based models utilizing EHRs are also the framework to solve other problems. BEHRT was integrated with observational causal inference techniques to estimate treatment effects, such as risk ratios [38]. Patient representations learned from a modified CEHR-BERT were used to study the synthesis of EHRs in [39]. Foresight [40] forecasts temporally-aware future clinical concepts within a patient trajectory leveraging structured and unstructured EHRs.

Recent EHR modeling works have shifted towards architectures beyond BERT, often surpassing it at longer context lengths [34, 22, 25, 26]. Antikainen et al. [34] demonstrated that XL-Net [41] is a competitive alternative to BERT capturing more positive cases. TransformEHR [22] introduced a new pretraining objective by forecasting all diagnoses in a single visit using an encoder–decoder framework. Fallahpour et al. [25] demonstrated that Mamba paired with multi-task learning outperformed Transformer-based architectures for a context length of 2048. However, the comparison did not investigate more modern Transformer architectures or the Mamba performance at varying context lengths. Wornow et al. [26] evaluated sequence models on various benchmark prediction tasks and EHR properties, such as token repetition and visit irregularity, showing that Mamba is more robust to longer context lengths beyond 4k tokens but achieves a reduced discriminative performance compared to Transformer-based models on 1k tokens. Given Mamba's competitive performance compared to Transformers, Mamba has found widespread applications in DL fields, including language models [42] or computer vision [43]. However, it is still unexplored how Mamba performs on shorter sequences (≤ 512 tokens), in clinical settings especially.

The following ablation studies were conducted in EHRs modeling: BRLTM [20] removed clinical concepts sequentially during pre-training and found that the precision increased towards smaller vocabularies and larger embedding size, whereas the size of the optimized model hyperparameters decreased. Different patient representations were compared in CEHR-BERT [21] highlighting the importance of the [ATT] token, whereas Wornow et al. [26] findings indicate that the model-dependent benefits of [ATT]. Moreover, Wornow et al. [26] compared various recent architectures at longer context lengths, ranging from 512 to 16384 tokens conditioned on the model architecture. The utilization of complete diagnosis and medication codes was reported as minor gains in CORE-BEHRT [35]. Therefore, this work fills a gap for an ablation study which provides insights from various properties, covering tokenization, model configurations, temporal data preprocessing and data availability, for different model architectures.

3. Methods

3.1. Heart failure dataset

The patient data was extracted from a clinical data warehouse (CDW) built from EHRs covering all emergency in-hospital and public specialized out-patient care within the Region Västra Götaland (VGR) of Sweden (total population 1.8 million). The study conforms to the principles outlined in the Declaration of Helsinki. The study was approved and informed consent from the study participants was waived by the Swedish Ethical Review Authority, through the Ethics Committee of Umeå University (EPN reference: DNR 2021-02786). Section 3.2 defines the clinical prediction tasks and subsequently presents the descriptive statistics of the cohort. Supplementary Table S1 shows the collected concepts of the patient data including demographics (age, sex, body mass index), diagnoses (DX), vital signs (VIT), laboratories (LAB), medications (MED) and procedures (PRO). While diagnoses and medications are assessed by international

classification systems, International Classification of Diseases 10th revision (ICD-10) and Anatomical Therapeutic Chemical (ATC) system respectively, procedures are classified based on the Swedish Classification of Care Measures (KVÅ[1]) system. To investigate the performance of sequence models in severe HF, a cohort was defined consisting of patients ≥ 18 years old at their first in-hospital HF diagnosis (Dx), identified by qualifying ICD-10 codes (I110, I130, I132, any I42 or any I50 in any position). The study period spanned between January 1, 2015 and December 31, 2022, resulting in $N = 42820$ HF patients. To include one-year prediction outcomes for all patients, the observation period ended December 31, 2023.

3.2. Clinical prediction tasks

Three clinical binary classification tasks were evaluated using the HF cohort at two different clinical events, initial HF hospitalization and latest (most recent) hospitalization, defining two patient trajectories (Fig 2a). For each task, the prediction window was within one year after discharge from hospital. The first task T_1 predicted clinical instability, a phenotype of unplanned readmission to hospital, following the initial HF Dx (prevalence $P_{T_1} = 39.7\%$), the second task T_2 predicted mortality following the initial HF Dx ($P_{T_2} = 24.8\%$), and the third task predicted mortality following the latest hospitalization ($P_{T_3} = 46.7\%$). Clinical instability was defined as being rehospitalized for a period of at least 48 hours that met one of the following criteria (Supplementary Table S2): i) admission to a clinical service unit relevant to HF care (e.g., cardiology or internal medicine), or ii) admission in a surgical service unit where a patient underwent a surgical procedure relevant to HF determined by selected procedure codes (e.g., an implantation of a cardiac device or heart valve replacement). The minimum duration was set to 48 hours to exclude hospitalizations which were less severe or possibly planned (e.g., routine checkups).

The demographic profile (Table A.1) of the HF cohort ($N = 42820$, 45.4% women) was median 80 (interquartile range (IQR) 72, 87) years old at the initial HF admission and 82 (73, 88) at the latest. In comparison, the initial HF admissions occurred earlier in a patient's life covering less historical events, resulting in shorter health trajectories (153 (71, 324) vs. 369 (170, 740) tokens) and fewer hospitalizations (2 (1, 3) vs. 4 (2, 6) admissions). Both trajectories share 7178 patients (16.8%) who were hospitalized exactly once.

3.3. Patient sequence representation

For outcome prediction, patient-specific EHRs were chronologically combined into a single sequence, with the latest clinical events occurring right-sided. Only hospitalizations with at least one diagnosis were included, excluding out-patient encounters that were assumed to indicate low clinical severity. Consecutive hospitalizations less than 24 hours apart were merged into one. Any hospitalization that led to in-hospital death was excluded from the input sequence, as being irrelevant to the outcomes. To train the sequence models, the events from all patient sequences were tokenized and mapped to a shared baseline vocabulary. ICD-10 and KVÅ were registered at the third hierarchical system-specific level. Related KVÅ codes with a digit in the third position were standardized by replacing the digit with an uncommon shared character, whereas alphabetic codes remained unchanged. MED categories were constructed by concatenation of the ATC code (level 4), the route of administration (e.g., intravenous or oral) and the amount of medication but not the dosage intervals. All VIT, LAB and MED concepts were uniformly discretized using 10 bins. Missing body mass index values were median imputed.

Fig. 2b shows the embeddings of an example patient sequence. A patient sequence contains up to v time-irregular visits. Each visit contains a varying number of clinical concepts (DX, VIT, LAB, MED, PRO). A single patient sequence is represented as an embedding vector $E^{1\times C\times d_m}$, where the context length C determines the maximal length of the right-sided sequence (look-back cutoff window) and d_m the hidden size of the embedding vector (internal representation). Events earlier than the specified C are discarded. The embeddings are an aggregation of the same-sized individual embeddings of the clinical concepts and token types to contextualize the events, demographic triplet (age, sex, body mass index), the passed time in weeks given a reference date, visit number to determine the respective hospitalization, visit segments to differentiate between two consecutive hospitalizations and absolute position within the sequence. Following previous works [21, 25], tokenization uses standard special tokens like [CLS] and [PAD] to mark sequence boundaries, as well as visit-specific tokens [VS] and [VE] for encounter start and end, and the artificial time token [ATT] (e.g., [M2] for a two-month gap) to indicate temporal intervals between consecutive encounters. The register token [REG] can be interpreted as a storage for global information after each encounter [44].

3.4. Implementation

This work compared three architecture classes used for DL sequence models: (i) Transformers (BERT [18], XLNet [41]), (ii) Transformers++ (ModernBERT [31] abbreviated as MBERT, Llama [23]), and (iii) Mambas (Mamba [24], Mamba2 [32]). The respective architectural design choices are described in Section Appendix B.1 and a theoretical comparison of Transformer and Mamba is described in Section Appendix B.2. To compare with traditional machine learning (ML), the commonly used eXtreme gradient boosting machine (XGBoost) [45] was included.

For each clinical task, the data was divided into a development (85%) and testing (15%) set stratified by the outcome of interest. The development set was further split into a training (90%) and validation set (10%). No patient appeared in more than one set. Here, the development set was used for both pre-training the sequence model on a model-specific learning objective and fine-tuning on the relevant clinical task. Importantly, patients were consistently assigned to the same set and split across both stages. We motivate this choice by our pre-selection of only HF patients as opposed to other, larger cohorts with more general clinical profiles [19, 29, 21, 33].

[1] https://www.socialstyrelsen.se/statistik-och-data/klassifikationer-och-koder/kva/

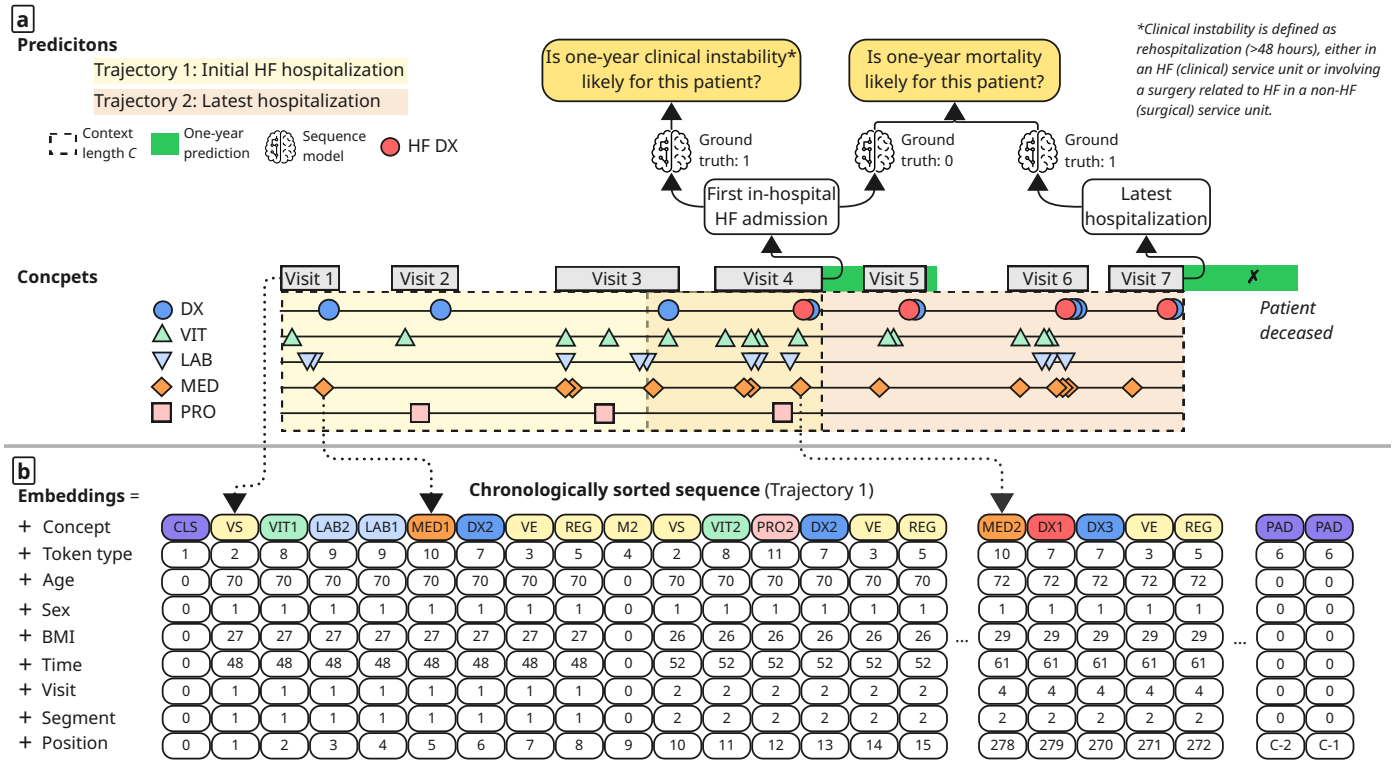


Fig. 2: (**a**) Overview of the clinical prediction tasks in this study: Given a simplified chronologically sorted patient sequence, separate sequence models were trained to predict at discharge: one-year clinical instability at the initial HF diagnosis in-hospital (trajectory 1), one-year mortality at the initial HF diagnosis in-hospital, and one-year mortality at the time of the latest hospitalization (trajectory 2). (**b**) Corresponding input sequence representation after the initial HF diagnosis trajectory highlighting the aggregated embeddings of tokenized concepts, concept token types, demographics (age, sex, body mass index (BMI)), time of events, visit numbers, alternating visit segments and absolute token positions.

Our code implementation is adapted from odyssey[2] [25]. All sequence models were trained using a single NVIDIA A100 80GB GPU on the same patient embeddings (Fig. 2b). Pre-training was set to 30 epochs with a patience of 5 epochs, fine-tuning to 10 with a patience of 2 epochs. In both stages, the optimizer choice was AdamW [46] with a learning rate of $5 \cdot 10^{-5}$ and a batch size of 32. In all dropout layers, the probability was set to 0.1. The masked language modeling (MLM) objective was masked with a probability of 0.15 in BERT and MBERT. For both Mambas, the size of the space state vector was set to 16, the size of the convolutional kernel to 4 and the Mamba block expansion to 2 matching the size of Transformer blocks [24]. XGBoost was trained using the default parameters (a total of 100 trees, a maximum depth of 6 per tree, learning rate of 0.3) on the frequency of clinical concepts within a patient sequence, excluding any temporal encounter information [21, 33, 25]. During fine-tuning, predicted probabilities were obtained through a two-layer feedforward classification head applied on the last hidden state of the rightmost non-padded token, and the backbone model was further trained using binary cross-entropy loss.

Due to the low prevalence of outcomes in the clinical HF cohort (Table A.1), e.g. mortality following the initial HF Dx has a prevalence of 24.8%, the area under the precision-recall curve (AUPRC) was reported as primary metric for the discriminative evaluation. Unlike the area under the receiver operating characteristic curve (AUROC), AUPRC captures better the model's ability to identify true positives in the presence of class imbalance. The calibration capability was assessed using the Brier score which measures the mean squared error between predicted probabilities and actual outcomes. All metrics were reported with 95% confidence intervals (CIs), computed via bootstrapping (1000 iterations).

3.5. Experiments

To investigate the performance on different clinical tasks for varying events and model architectures, this paper formulates the research questions (RQs) as follows:

- **RQ1** How does token granularity of the input sequence influence the prediction of patient outcomes?
- **RQ2** How do architectural changes, including variations in context length and model size, influence the predictive performance?

[2] https://github.com/VectorInstitute/odyssey

- **RQ3** How do different preprocessing techniques for constructing temporal sequences, namely truncation and aggregation, impact the representation of the patient history and the resulting model predictions compared to the sequence cutoff?
- **RQ4** How does the model performance scale with varying data availability of clinical concepts and training data?

The experiments conducted to address these RQs are outlined in Fig. 1:

3.5.1. Ablation of input patient sequence

To answer **RQ1**, the token granularity of the shared patient sequence vocabulary was studied in two ways. Finer encoded token granularity typically leads to a larger vocabulary. Firstly, the granularity of the discrete intervals of VIT, LAB and MED is affected by the chosen number of bins $b \in \{5, 10\}$ bins, representing 5 or 10 categories, respectively. Secondly, the hierarchical ICD-10 code level i is commonly set to the category code level ($i = 3$, e.g. I50) [33, 35]. However, the diagnosis level ($i = 4$, e.g. I509) can enrich the diagnostic profile through a more precise assessment, thereby allowing the identification of more detailed phenotypes within the HF cohort. Table B.3 presents the contributions of the concept-specific unique tokens to the total vocabulary for all combinations of b and i leading to varying vocabulary sizes.

3.5.2. Ablation of model configuration

To investigate **RQ2**, the model architecture was modified by the context length C and the architecture size S. The context length indicates how many tokens of an input patient sequence can be processed for a given model. Any concept of the sequence beyond the context length is excluded from the input sequence. Context lengths $C \in \{128, 256, 512\}$ were used throughout the architectural ablations. A longer C captures long-range dependencies in the sequence enabling improved contextualization but suffers from computational expenses, particularly in Transformers due to the quadratic memory usage of self-attention [9]. Despite the advantages with longer C in [25, 26], this work focused on $C \leq 512$ determined by the collected data and number of hospitalizations. The vocabulary $V_{b=10,i=3}$ was maintained across all C to eliminate variations unrelated to C. Table B.4 shows the effect of C on the initial HF and latest hospitalization, as patients at their latest admission had longer medical records, reflected by higher median tokens, more visits and greater longitudinal gap ΔDays between the last and first collected event, as also visualized in Supplementary Fig. S1.

To further analyze the performance efficiency, each model was compared across different architecture sizes $S \in \{\text{Tiny}, \text{Small}, \text{Medium}\}$, with configurations shown in Table B.5. The modifications of S were controlled by the dimension of hidden layers d_{m} and the size of the feed-forward layer d_f. For Transformer-based architectures, the number of hidden layers n_m and number of attention heads n_h were adjusted, while for Mamba-based architectures, the number of Mamba blocks n_b to ensure comparable total number of model parameters.

3.5.3. Ablation of patient's medical history

To address **RQ3**, the patient history H was modified by truncation and aggregation of the existing hospitalizations before tokenization. The vocabulary $V_{b=10,i=3}$ and $C = 512$ was used across all temporal modifications to avoid introducing additional confounding effects. Table B.6 highlights the data modifications through truncation, cutoff and aggregation in order of extended temporal horizons, as also visualized in Supplementary Fig. S2. Here, cutoff refers to the baseline right-sided patient history $H_{t \leq C}$ determined by C, where prior events beyond this cutoff are discarded. Truncation investigated how a predefined temporal look-back window affected the model performance, motivated by how much collected patient history is either necessary or obsolete to achieve a performance similar to $H_{t \leq C}$. Look-back windows included either only the concepts of the latest visit (discharged alive) or additionally up to $\{1, 3\}$ years of prior records. Contrary, aggregation extended the patient sequence by averaging reoccurring numerical concepts (e.g., ongoing monitoring of vital signs or medications) within a temporal window of $w \in \{1, 2\}$ hospitalization days. A longer w reduced the number of tokens within the input sequence but revealed more prior history compared to $H_{t \leq C}$.

3.5.4. Ablation of data availability

The performance variations across model types were simulated for different data availability (**RQ4**), namely variations in clinical concept groups and size of training data. Firstly, to understand how different architectures leverage additional concepts for predictive power, a baseline DX sequence was augmented incrementally with events from the added concept group without extending the temporal information. For example, the combination DX+VIT excluded all LAB+MED+PRO tokens and padded the sequence right-sided, without incorporating theoretically available historical information beyond C. The incremental order was DX, VIT, LAB, MED, PRO to reflect the relative concept accessibility within EHRs and resource cost. This ordering assumed that DX and VIT are routinely recorded during care at low-cost. Contrary, LAB, MED and PRO rely on the integration from external data sources and are at high-cost. Secondly, varying sizes of the training data simulated the performance and generalization capacity of sequence models in clinical settings with limited data availability.

4. Results

4.1. Ablation studies

Unless otherwise specified, all models were trained with vocabulary $V_{b=10,i=3}$, context length $C = 512$ and model size $S =$ Medium across all clinical prediction tasks. Although related models, such as BEHRT or Med-BERT, are not directly applicable anymore due to differences in the embedded patient representation and clinical concepts, the BERT architecture was adopted as baseline reference due to its established effectiveness in EHR sequence modeling. For all ablation studies, the visualizations of AUROC and Brier score, as well as the numerical results of all metrics (including AUPRC), are reported in Supplementary Section S3.

4.1.1. Impact of vocabulary modification (**RQ1**)

Fig. 3 visualizes the AUPRC performance for four different token vocabularies. The vocabularies are sorted in ascending order of unique tokens determined by the number of bins b and hierarchical ICD-10 code level i (Table B.3). Overall, Llama achieves the best discriminative performance (highest AUPRC and AUROC) and best calibration capability (lowest Brier score) across nearly all combinations, although Mamba2 is competitive for T_2. For Llama, finer discretized VIT, LAB and MED measurements ($V_{b=10,i=3}$) perform best at the initial trajectory, while finer diagnoses ($V_{b=5,i=4}$) perform best at the recent trajectory. A similar behavior can be observed for Mamba2, except for T_1. BERT, XLNet and MBERT exhibit no generalizable trend, with XLNet being the most sensitive model to any token modification. These findings highlight that fine-grained measurement resolutions are favorable design choices at the initial trajectory, while fine-grained DX tokens yield improvements at the latest trajectory for Llama and Mamba2. Moreover, models trained on next token prediction (NTP) outperform models trained otherwise in both predictive performance and calibration capability, with Llama demonstrating superiority across most task and vocabulary ablations.

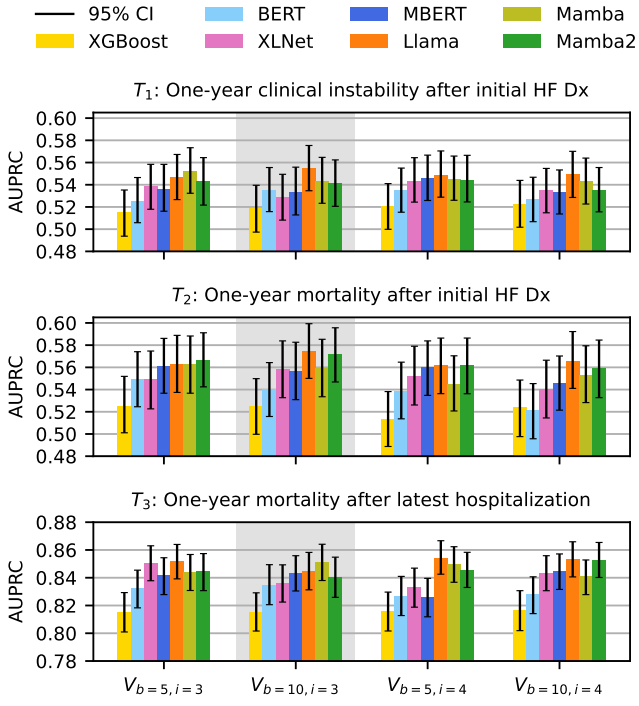


Fig. 3: Ablation of the vocabulary V across three clinical tasks T evaluated by bootstrapped AUPRC (↑) for Medium-sized sequence models and $C = 512$. The resolution of the vocabulary increases from lowest (left) to highest (right). Gray background highlights common setup shared across all ablations.

4.1.2. Impact of context length and architecture size (**RQ2**)

Fig. 4 presents the AUPRC results for different configurations of context lengths $C \in \{128, 256, 512\}$ and architecture sizes $S \in \{\text{Tiny}, \text{Small}, \text{Medium}\}$ (abbreviated as T, S, M). The configurations are sorted from lowest to highest complexity represented by context length (Table B.4) and model size (Table B.5). At $C = 512$, the best AUPRC is achieved by M-Llama in tasks T_1 and T_2 and S-Mamba in T_3, followed marginally by S-Llama. Most notably, T-Llama at only $C = 128$ demonstrates superiority over the best performing BERT/XLNet/MBERT of any C (except M-MBERT at $C = 256$) throughout all tasks. Additionally, T-Llama outperforms the best Mambas for $C \leq 256$, aside M-Mambas for mortality predictions. T-Mambas highlight similar observations but are more often outperformed by M-MBERT (all tasks) or M-XLNet (T_2). In nearly all tasks, T-Llama and T-Mambas at $C = 256$ surpass the respective best configurations at $C = 128$, linking the learning capability rather to C than S. Extending C for S-Llama and S-Mambas at $C = 256$ is superior to the at most marginal gain seen with M-variants suggesting overfitting towards larger sizes. Similarly, S-Mambas are sufficient choices at $C = 512$. MBERT indicates scalability issues or parameter inefficiency in smaller configurations as only M-MBERT outperforms BERT and XLNet consistently. Supplementary Table S3 shows that $C = 1024$ boosts the performance for M-variants of Llama and Mambas in T_3 which further underlines the capability of the respective models to capture long-range contextual interactions. For this cohort and prediction tasks, these results suggest that performance gains stem primarily from data quantity (extended C) as opposed to enriched representation capacity (increased S). More specifically, S-Llama delivers a strong performance that is robust for any C across all tasks.

4.1.3. Impact of patient's medical history (**RQ3**)

Fig. 5 visualizes the AUPRC outcomes for the temporally modified patient records. The patient history is sorted in ascending order of temporal resolution, ranging from predefined look-back time windows modified by truncation to summarized clinical events modified by aggregation (Table B.6). Across all tasks and nearly all temporal modifications Llama demonstrates strongest predictive power. Overall, truncation and aggregation techniques lead to performance gains over or match the performance of the baseline cutoff $H_{t\leq C}$ across all models and tasks, except for Mamba2 in T_2. Truncations perform mostly the best in BERT ($H_{t\leq 1\text{y}}$) and MBERT ($H_{t\leq 3\text{y}}$) with marginal gains over aggregations. XLNet benefits from a weak aggregation ($H_{t\leq C,w=1\text{d}}$). For Mambas, truncation typically results in a performance drop, and aggregation is only beneficial in Mamba, except for Mamba2 in T_3 where both help. Llama matches the performance of $H_{t\leq C}$ for mortality predictions with the available information of the recent hospitalization ($H_{t=0}$) and outperforms its performance by aggregation eventually, whereas $H_{t\leq 1\text{y}}$ matches $H_{t\leq C}$ in T_1. These outcomes indicate that a history cutoff through C is a suboptimal design choice. However, the temporal modifications showcase model-specific improvements or degradations and limited capability for generalization. Aggregations, which trade increased temporal resolution for a lower proportion of tokens, show a promising direction to enhance the ability to learn from in Llama and Mamba.

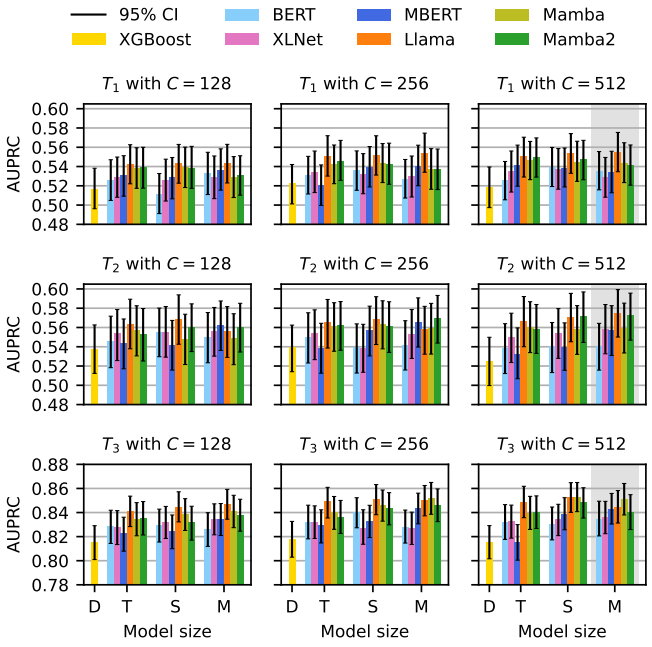


Fig. 4: Ablation of the context length C and model size evaluated by bootstrapped AUPRC (↑) across three different clinical tasks. Within each C sequence modes are sorted by Tiny, Small, and Medium configuration, and compared to XGBoost's Default configuration. Gray background highlights common setup shared across all ablations.

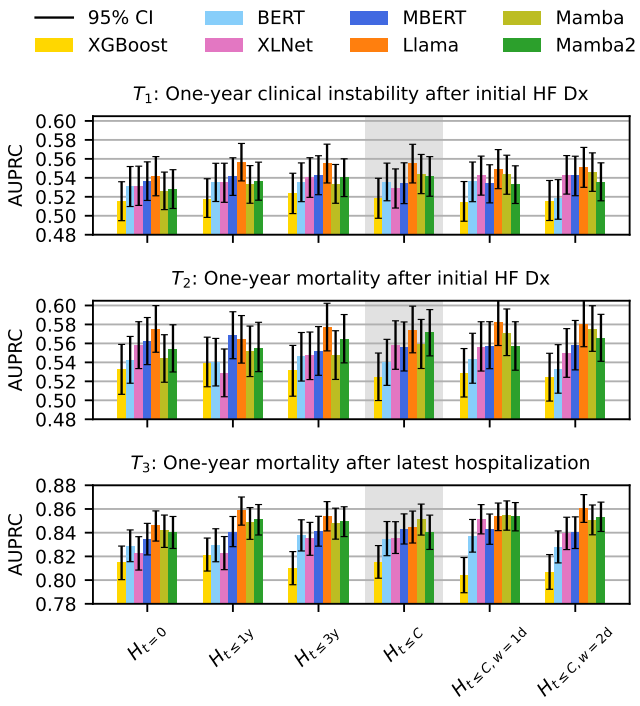


Fig. 5: Ablation of the patient history H: The historical record is extended from the latest available hospitalization (left) to a prolonged context (right). All modifications are evaluated by bootstrapped AUPRC (↑) using Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.

4.2. *Data scalability analysis*

All experiments analyzing the scalability of models in different data settings were trained with a context length $C = 512$ and model size S = Medium. For the incremental concept augmentation, a concept-specific vocabulary was extracted from $V_{b=10,i=3}$ and used for training, whereas the training size variation experiment employed $V_{b=10,i=3}$.

4.2.1. *Impact of incremental concept augmentation (**RQ4**)*

Fig. 6 compares the AUPRC outcomes for the concept augmentation, incrementally introducing a new clinical concept starting from only DX until all collected concepts are included. Llama consistently demonstrates the best discriminative capability across almost all task and concept combinations. The inclusion of all concepts benefits each model only in T_2, whereas DX+VIT+LAB+MED tends to perform best in T_1 and T_3. Thus, the augmentation by PRO adds primarily noise to the predictability in T_1 and T_3. Unlike Mamba, Mamba2 potentially capitalizes multiple concepts better in mortality prediction tasks (DX+VIT+LAB+MED vs. DX+VIT) making Mamba more viable in cases with fewer available concepts.

4.2.2. *Impact of training set size (**RQ4**)*

The impact of different training size ratios $\{25\%, 50\%, 75\%, 100\%\}$ evaluated by AUPRC on the same-sized test set is presented in Fig. 7. Most sequence models trained on 50% data outperform XGBoost trained on 100%. For nearly all tasks, Llama trained on 75% data surpasses other models trained on 100%. In mortality prediction tasks, Llama converges between 75% and 100% training data, highlighting that less data is sufficient to reach its capacity. Similarly, Mamba2 hits a plateau from 50% to 100% in T_1 and T_3, with initial gains from 25% to 50%. Contrary, Mamba benefits from the full training set, suggesting a slower convergence. These results show that Llama achieves nearly competitive performance with the full training set when utilizing 25% less data, suggesting greater training efficiency through more effective representation learning. Additionally, Llama demonstrates robust scalability across both scenarios of low- and high-data availability scenarios.

5. Discussion

Accurately predicting clinical instability and mortality, within one-year from the initial HF hospitalization in particular, are prioritized tasks given the rapid progression and severity of HF. So far, most studies have designed prediction models based on general patient cohorts [19, 35, 25, 26] or fine-tuned a general cohort for disease-specific subcohorts [20, 29, 33]. To reflect real-world clinical practice, however, there is potentially a greater demand for AI models tailored around subpopulations, as specific diseases, such as HF, are associated with distinct

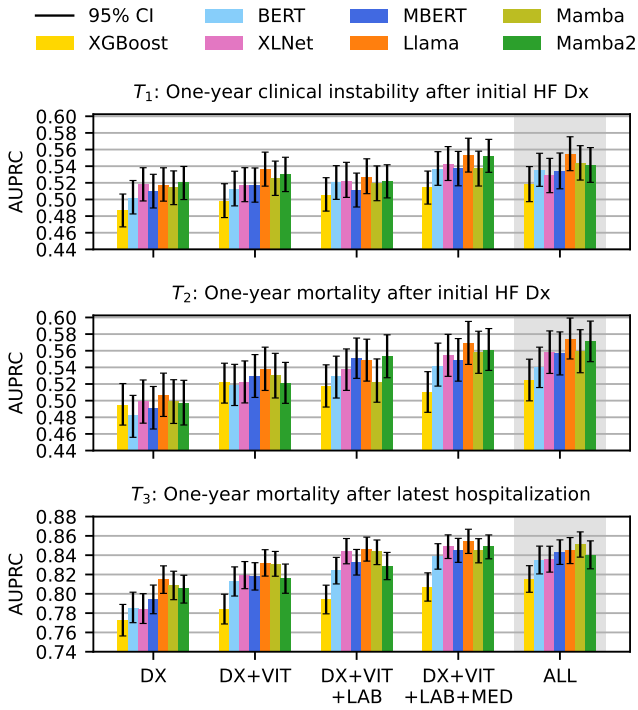


Fig. 6: Augmentation of DX through incremental addition of the respective clinical concept, continuing until all concepts including PRO are selected. Each augmentation is evaluated by bootstrapped AUPRC (↑) using Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.

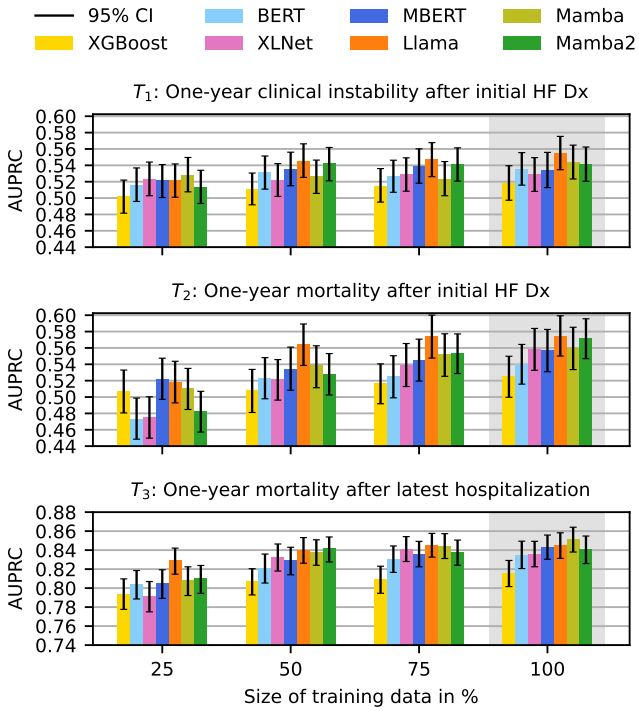


Fig. 7: Ablation of different training sizes for Medium-sized sequence models and $C = 512$ evaluated by bootstrapped AUPRC (↑). Gray background highlights common setup shared across all ablations.

clinical concepts. Moreover, the accessibility and availability of large, general-purpose cohorts for model development cannot be always guaranteed. Consequently, the utilization of a subpopulation for prediction tasks is more realistic as well as clinically meaningful.

Despite the highly successful application of BERT for EHR-based predictions, up-to-date architectures, such as Transformers++ and Mambas, have not yet been fully explored for shorter context lengths ($\leq$ 512 tokens), especially across various settings. This setting is realistic for many applications involving EHR or hospitalizations, as their clinical events are typically sparse and infrequently recorded. To investigate such settings, we analyzed empirically how token modifications of the input sequence (**RQ1**), architectural design choices (**RQ2**), and temporal preprocessing techniques of the patient history (**RQ3**) affected the discriminative performance and calibration capability of six sequence models across three model types.

Overall, our results demonstrate that Llama exceeds Mamba-based architectures, which in turn outperform the commonly used BERT (baseline indicator), in discrimination, calibration and robustness between and within nearly all ablations and prediction tasks. The findings highlight that Llama is more effective than Mamba for the shorter context lengths examined in our study. However, long context lengths were not fully evaluated, although Mamba is expected to excel with longer contexts [24]. Furthermore, models that are trained on the NTP objective achieve better results compared to models trained on MLM. Given the promising performance of Llama and Mambas compared to other models, we focus our discussion primarily around these.

The ablation of the token granularity (Fig. 3) suggests that performance improvements depend on model candidate and trajectory position, highlighting poor generalization capability, in BERT/XLNet/MBERT especially. Interestingly, the most fine-grained vocabulary $V_{b=10,i=4}$ with enriched measurement and diagnosis representations performs the worst in all tasks. Llama and Mamba2 tend to empower a richer discretization of measurements at the initial HF Dx and instead detailed diagnosis codes at the latest hospitalization. At the initial HF Dx the severity of the condition is less pronounced and potentially more influenced by variations in the measurement tokens, whereas during progression the Dx profile becomes more defined leveraging precise diagnoses at the latest trajectory.

Both Llama and Mambas are more efficient at representation learning as their respective T-variants mostly outperform the best BERT/XLNet/MBERT setup of higher complexity for the same context length C throughout all tasks (Fig. 4). The S-variants of Llama and Mambas are the most effective studied configurations, prioritizing computational efficiency and stability. Smaller models are more stable due to their reduced complexity, which minimizes the risk of overfitting to noise in the training data. Moreover, both architecture types leverage contextual informa-

tion over model complexity to improve performance. Noticeably, Mamba has the lowest number of model parameters compared to Mamba2 and Llama (Table B.5).

In the temporal modifications of the patient history, cutoff ($H_{t\leq C}$) assumes relevance for each measured or assigned clinical concept and of recent activity given a right-sided context length. However, cutoff discards prior history given the fixed context length or may introduce noise due to overrepresentation. Truncation and aggregation tested the assumed relevance by evaluating their performance effects in relation to cutoff. Truncation clipped token sequences to the most recent hospitalizations systematically while reducing the number of concepts. In contrast, aggregation compressed high-amount token sequences into semantically meaningful and temporally enriched summarizations. Whereas a cutoff history peaks only in Mamba2, Llama and Mamba tend to yield better performances for aggregated sequences (Fig. 5). Hence, aggregation challenges the assumption that an uncompressed cutoff sequence is the best design choice for Llama and Mamba.

The scalability analysis reveals that 75% training data of all concepts outperforms 100% training data limited to DX+VIT+LAB concepts, highlighting the predictive power of concept diversity over cohort size (Fig 6 and Fig 7). However, the inclusion of PRO mostly caused deterioration. One possible explanation is that the mapping of procedure codes (KVÅ) has a more limited hierarchical construction, failing to account for order of levels in taxonomy. An alternative could be to determine semantic similarity based on the code descriptions, using a predefined threshold. The efficiency of Llama and Mambas ($C = 512$, M-variant) becomes further evident as both exhibit superiority on 25% less training data compared to other models trained on the complete data (Fig 7).

There are several limitations in our work. Firstly, our analysis is restricted to a Swedish HF cohort and lacks external validation. Although HF patients represent a heterogeneous population, our findings may not generalize to other disease cohorts. Secondly, the laboratories and medications were selectively curated for HF, which restricts information and potentially limits the models to leverage learnable out-of-domain knowledge. Moreover, alternative formulations of the patient embeddings, such as the pre-concatenation of demographic information into the sequence instead of assigned embeddings, were not explored. Future work should focus on fusing the multi-modal nature of EHRs, such as cardiac image data or parameters (e.g., left ventricular ejection fraction) and clinical text notes, with the token sequences to explore performance benefits. Additionally, the inclusion of information for specific medical disciplines and out-of-hospital settings provided by primary care or out-patient visits could prospectively enrich patient embeddings and trajectories.

Lastly, based upon our findings we recommend design principles as an initial guide for future works in EHR sequence modeling with HF patients. This recommendation encapsulates the most important findings for model development, reflecting predominant trends of the ablation studies rather than task-specific best performing configurations. The Llama architecture is clearly recommended due to its consistently highest performance across all analyzed experiments for shorter context lengths (≤ 512 tokens). Specifically, a small-sized vocabulary ($V_{b=10,i=3}$) with fine measurements and simplified ICD-10 codes at category level covers sufficient token granularity. Our findings indicate that model configurations should prioritize context length over model complexity. Therefore, a small-scaled Llama with a context length $C = 512$ is suggested. Moreover, a model of lower complexity with extended context length is a viable choice in a clinical environment as it maintains strong predictive performance while remaining computationally efficient. Among the temporal preprocessing techniques, an aggregation of events is recommended instead of truncation due to the greater performance improvements. However, further analysis is required to fully capture potential benefits of semantically and temporally compressed sequences.

6. Conclusion

Sequence models using EHRs have gained popularity due to their superiority across various clinical prediction tasks. However, systematic performance insights on varying design choices are limited. In this paper, we present a first ablation study to investigate the behavior of sequence models applied to a HF cohort, examining the impact of modified tokenizations, model architectures and temporal preprocessing techniques. Our findings show that Llama, followed by Mambas, consistently outperforms the widely used BERT in all tasks and ablations further indicate that Llama scales and trains more efficiently. These insights offer guidance for future clinical model development, targeting more efficient and effective design decisions regarding technical setup and empirical study design in EHR-based sequence modeling.

CRediT authorship contribution statement

Falk Dippel: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yinan Yu:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Annika Rosengren:** Writing – review & editing, Resources, Conceptualization. **Martin Lindgren:** Writing – review & editing, Conceptualization. **Christina E. Lundberg:** Writing – review & editing. **Erik Aerts:** Writing – review & editing. **Martin Adiels:** Writing – review & editing, Formal analysis, Conceptualization. **Helen Sjöland:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no competing interest.

Acknowledgments

This work was supported by grants from the Swedish state under an agreement concerning research and education of doctors (ALFGBG-991470 [to H.S.]), the Swedish Research Council (VRREG 2023-06421 [to H.S.]), Forte – the Swedish Research Council for Health, Working Life and Welfare (DNR 2024-00277 [to H.S.]), and Vinnova Advanced Digitalization (DNR 2024-01446 [to Y.Y. and H.S.]). The authors would like to thank the VGR AI platform for providing the essential GPU resources to conduct the experiments.

Data availability

The data underlying this paper are not available for sharing due to participant privacy and ethical restrictions.

Code availability

The code is not publicly available but will be made available upon reasonable request reviewed by the authors. The following packages were used: torch 2.2.0, transformers 4.49.0, mamba-ssm 2.2.2, causal-conv1d 1.4.0, and xgboost 2.0.0.

Appendix A. Data

Table A.1 presents the statistical properties of the HF cohort.

Table A.1: Statistical description of the HF cohort given patients' trajectory and prevalence of clinical prediction task.

Information	Initial HF Dx (Trajectory 1)	Latest HF Dx (Trajectory 2)
N	42820	42820
Age (right-sided)	IQR: 80 (72, 87), min: 18, max: 107	IQR: 82 (73, 88), min: 18, max: 107
Sex (% women)	45.4%	45.4%
Hospitalizations	IQR: 2 (1, 3), min: 1, max: 61	IQR: 4 (2, 6), min: 1, max: 80
Longitudinal gap first to last event	IQR: 32 [6, 741], min: 0, max: 3075	IQR: 783 [127, 1634], min: 0, max: 3578
All events	IQR: 153 (71, 324), min: 1, max: 9552	IQR: 369 (170, 740), min: 2, max: 10510
DX events	IQR: 9 (5, 16), min: 1, max: 311	IQR: 20 (10, 38), min: 1, max: 881
VIT events	IQR: 30 (0, 91), min: 0, max: 2634	IQR: 98 (27, 227), min: 0, max: 3433
LAB events	IQR: 23 (8, 55), min: 0, max: 3845	IQR: 54 (22, 123), min: 0, max: 5121
MED events	IQR: 63 (26, 137), min: 0, max: 5746	IQR: 142 (59, 309), min: 0, max: 6236
PRO events	IQR: 3 (1, 7), min: 0, max: 339	IQR: 7 (3, 13), min: 0, max: 350
Prevalence of one-year mortality	24.8%	46.7%
Prevalence of one-year instability	39.7%	-

Appendix B. Methods

Appendix B.1. Sequence model architectures

This section compares the different architectural design choices while Table B.2 summarizes the design choices.

Table B.2: Comparison of architectural key design choices. MBERT=ModernBERT.

Model	Learning objective	Position encodings	Activation function	Applied normalization
BERT	MLM	Absolute	GeLU	LayerNorm (post)
XLNet	PLM	Relative	GeLU	LayerNorm (post)
MBERT	MLM	RoPE	GeGLU	LayerNorm (pre)
Llama	NTP	RoPE	SwiGLU	RMSNorm (pre)
Mamba	NTP	-	SiLU	RMSNorm (pre)
Mamba2	NTP	-	SiLU	RMSNorm (pre)

BERT captures bidirectional contextual information via self-attention and is pre-trained on a MLM objective, where tokens of the input sequence are randomly masked and predicted given the surrounding context. XLNet predicts tokens autoregressively utilizing a permutation language modeling (PLM) objective. In PLM, tokens are randomly permuted based on a factorization order allowing to capture more long-range dependencies beyond the fixed context provided by auto-encoders [41]. Although PLM undermines the temporal structure of EHR sequences, this flexibility can be advantageous when capturing interactions among co-occurring clinical events. ModernBERT (hereafter referred to as MBERT) and Llama include several improvements over BERT. The absolute position embedding is replaced with rotary position embedding (RoPE) [47] capturing both the absolute position of tokens and the relative distance between tokens through rotation encoding. MBERT replaces the Gaussian error linear unit (GeLU) activation function with Gaussian error gated linear unit (GeGLU) and Llama with a combination of GLU and Swish activation (SwiGLU) due to the apparent performance gains [48]. The post-normalization of the input after passing through a Transformer sub-layer is replaced with pre-normalization to improve training stability [49]. Llama replaces LayerNorm with root mean square layer normalization (RMSNorm) [50] reducing the computational burden by eliminating the mean calculation required in LayerNorm. For better representation learning MBERT introduces an alternating mechanism between global and local attention every three layers [31].

Mambas like Llama are pre-trained on the NTP objective, where the next token is predicted based on only past tokens. Unlike Transformers where the attention mechanism causes quadratic training time, the Mamba architecture combines discretized SSM with a selection mechanism that parameterizes the SSM to learn relevant context while hardware-aware algorithm enable linear scaling with sequence length [24]. Mamba2 improves Mamba's efficiency by introducing the state space duality which connects SSM with attention heads and simplifies the structure of the matrix $\mathbf{A}$ in the state equation of SSM into a scalar multiple of the identity [32]. Both Mambas utilize a sigmoid linear unit (SiLU) activation function and RMSNorm as pre-normalization.

Appendix B.2. Model Theory

This section explains the theoretical concepts behind the attention mechanism in Transformer and SSM modeling in Mamba.

Appendix B.2.1. Multihead attention in Transformer

To calculate the attention weights, a scaled dot-product attention is formulated as

$$A = \text{Softmax}\left(\frac{QK^\intercal}{\sqrt{d_k}}\right)V \tag{B.1}$$

where Q, K, V denote the query, key and value matrices and d_k the dimension of keys [9]. The vectors Q, K, V are linear projections of the input sequence x using weights as $Q = xW^Q$, $K = xW^K$, $V = xW^V$ where the weights $W^Q \in \mathbb{R}^{d_\text{m} \times d_k}$, $W^K \in \mathbb{R}^{d_\text{m} \times d_k}$ and $W^V \in \mathbb{R}^{d_\text{m} \times d_v}$ are trainable model parameters with d_m being the dimension of the embeddings and d_v being the dimension of value. The MHA weights are calculated as

$$\text{MHA} = (A_i \oplus \ldots \oplus A_{n_h})W^O \tag{B.2}$$

by concatenating the outputs of n_h parallelized attention heads, followed by a linear projection with the trainable output matrix $W^O \in \mathbb{R}^{n_h d_v \times d_\text{model}}$.

Appendix B.2.2. State space model in Mamba

The Mamba architecture is based on a discrete SSM formulation paired with convolutional computation power. In general, a continuous SSM describes the system dynamics through a state equation

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t) \tag{B.3}$$

using n-dimensional hidden states $h(t) \in \mathbb{R}^n$ and maps a state to an output $y(t) \in \mathbb{R}$ through the observation equation

$$y(t) = \mathbf{C}h(t) + \mathbf{D}x(t) \tag{B.4}$$

for a given input $x(t) \in \mathbb{R}$ at current time t using system matrices $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times 1}$, $\mathbf{C} \in \mathbb{R}^{1 \times n}$, $\mathbf{D} \in \mathbb{R}$ with the common choice of $\mathbf{D} = 0$. Given a discrete time step k, the discrete SSM is expressed as

$$h_k = \bar{\mathbf{A}}h_{k-1} + \bar{\mathbf{B}}x_k \tag{B.5}$$

$$y_k = \bar{\mathbf{C}}h_k \tag{B.6}$$

with the discretised matrix versions $\bar{\mathbf{A}} = \exp(\Delta\mathbf{A})$ and $\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I})$ obtained from zero-order hold discretization over a constant interval $\Delta = [t_{k-1}, t_k]$ [24]. Similar to multihead attention, the matrix calculation is formulated using a global convolution kernel $\bar{K}$

$$y = x * \bar{\mathbf{K}} \tag{B.7}$$

with $\bar{\mathbf{K}} = (\mathbf{C}\bar{\mathbf{B}}, \ldots, \mathbf{C}\bar{\mathbf{A}}^k\bar{\mathbf{B}}, \ldots)$.

Appendix B.3. Experiments

Tables B.3–B.6 highlight the varying properties of the patient sequences given the respective conducted ablation.

Table B.3: Ablation of vocabulary V: The vocabulary size is influenced by the number of bins b to represent VIT, LAB and MED, as well as if DX are represented by category ($i = 3$) or diagnosis ICD-10 codes ($i = 4$).

Vocabulary	Total	DX	VIT	LAB	MED	PRO
Patient trajectory up to initial HF hospitalization						
$V_{b=5,i=3}$	3293	1343	36	108	858	948
$V_{b=10,i=3}$	4128	1343	66	198	1573	948
$V_{b=5,i=4}$	6450	4500	36	108	858	948
$V_{b=10,i=4}$	7285	4500	66	198	1573	948
Patient trajectory up to latest hospitalization						
$V_{b=5,i=3}$	3565	1416	36	108	942	1063
$V_{b=10,i=3}$	4470	1416	66	198	1727	1063
$V_{b=5,i=4}$	7323	5174	36	108	942	1063
$V_{b=10,i=4}$	8228	5174	66	198	1727	1063

Table B.4: Ablation of context length C: The combination of C and patient trajectory influences the number of tokens, the number of hospitalizations (visits) and longitudinal gap ΔDays between last and first event.

Vocabulary	C	Tokens	Visits	ΔDays
Patient trajectory up to initial HF hospitalization				
$V_{b=10,i=3}$	128	128 (71, 128)	1 (1, 2)	11 (5, 46)
$V_{b=10,i=3}$	256	153 (71, 256)	1 (1, 2)	19 (6, 280)
$V_{b=10,i=3}$	512	153 (71, 324)	1 (1, 3)	27 (6, 542)
Patient trajectory up to latest hospitalization				
$V_{b=10,i=3}$	128	128 (128, 128)	1 (1, 2)	23 (9, 192)
$V_{b=10,i=3}$	256	256 (170, 256)	2 (1, 3)	122 (21, 574)
$V_{b=10,i=3}$	512	369 (170, 512)	3 (2, 4)	356 (54, 1030)

Table B.5: Ablation of model size S: All model parameters are listed in the order of Tiny/Small/Medium configurations for a vocabulary of 4470 tokens. The dimension of the hidden layers is denoted as d_m, the number of hidden layers as n_m, the number of (attention) heads n_h, and the number of Mamba blocks as n_b. The dimension of the feed-forward layer d_f was set to 512/1024/2048. In Mamba2, the dimension of heads d_p was set to 32/64/64.

Model	d_m	n_m	n_h	n_b	Parameters in M
BERT	256/512/512	4/4/6	4/4/8	-	2.4/11.2/21.7
XLNet	256/512/512	4/4/6	4/4/8	-	2.3/11.8/22.8
MBERT	256/512/512	4/4/6	4/4/8	-	2.5/13.1/27.7
Llama	256/512/512	4/4/6	4/4/8	-	3.6/15.1/29.8
Mamba	256/512/512	-	-	2/6/12	2.0/12.5/22.6
Mamba2	256/512/512	-	16/16/16	2/6/12	3.2/14.9/25.2

Appendix C. Supplementary data

Supplementary material related to this article can be found attached.

Glossary

References

[1] A. Groenewegen, F. H. Rutten, A. Mosterd, A. W. Hoes, Epidemiology of heart failure, European journal of heart

Table B.6: Ablation of patient's medical history H with context length $C = 512$: Truncation includes encounter information only within the latest admission ($t = 0$) or up to $t \in \{1, 3\}$ years of prior H. Cutoff refers to the unprocessed H determined by C. Aggregation averages continuous features within $w \in \{1, 2\}$ days during the encounter duration extending the longitudinal gap ΔDays between the first and last event within H.

Dataset	Modification	Tokens	Visits	ΔDays
Patient trajectory up to initial HF hospitalization				
$H_{t=0}$	Truncation	96 (50, 179)	1 (1, 1)	6 (3, 11)
$H_{t\leq 1y}$	Truncation	124 (61, 248)	1 (1, 2)	10 (4, 50)
$H_{t\leq 3y}$	Truncation	143 (68, 297)	1 (1, 2)	17 (5, 276)
$H_{t\leq C}$	Cutoff	153 (71, 324)	1 (1, 3)	27 (6, 542)
$H_{t\leq C,w=1d}$	Aggregation	112 (55, 229)	1 (1, 3)	31 (6, 647)
$H_{t\leq C,w=2d}$	Aggregation	81 (42, 161)	1 (1, 3)	32 (6, 704)
Patient trajectory up to latest hospitalization				
$H_{t=0}$	Truncation	103 (54, 188)	1 (1, 1)	5 (2, 10)
$H_{t\leq 1y}$	Truncation	215 (100, 436)	2 (1, 3)	46 (7, 184)
$H_{t\leq 3y}$	Truncation	311 (142, 512)	2 (1, 4)	201 (25, 591)
$H_{t\leq C}$	Cutoff	369 (170, 512)	3 (2, 4)	356 (54, 1030)
$H_{t\leq C,w=1d}$	Aggregation	262 (123, 512)	3 (2, 5)	510 (88, 1266)
$H_{t\leq C,w=2d}$	Aggregation	187 (91, 357)	3 (2, 6)	642 (116, 1429)

AI	Artificial intelligence
ATC	Anatomical Therapeutic Chemical
AUPRC	Area under the precision-recall curve
AUROC	Area under the receiver operating characteristic curve
BERT	Bidirectional encoder representations from Transformers
CDW	Clinical data warehouse
DL	Deep learning
Dx	Diagnosis
DX	Diagnoses
EHR	Electronic health record
HF	Heart failure
ICD-10	International Classification of Diseases 10th revision
KVÅ	Classification of Care Measures
LAB	Laboratories
LSTM	Long short-term memory
MBERT	ModernBERT
MED	Medications
ML	Machine learning
MLM	Masked language modeling
NLP	Natural language processing
NTP	Next token prediction
PLM	Permutation language modeling
PRO	Procedures
RNN	Recurrent neural network
SSM	State space model
VIT	Vital signs
XGBoost	EXtreme gradient boosting machine

failure 22 (8) (2020) 1342–1356.
URL https://doi.org/10.1002/ejhf.1858

[2] V. M. van Deursen, R. Urso, C. Laroche, K. Damman, U. Dahlström, L. Tavazzi, A. P. Maggioni, A. A. Voors, Co-morbidities in patients with heart failure: an analysis of the european heart failure pilot survey, European journal of heart failure 16 (1) (2014) 103–111.
URL https://doi.org/10.1002/ejhf.30

[3] N. Azad, G. Lemay, Management of chronic heart failure in the older population, Journal of geriatric cardiology: JGC 11 (4) (2014) 329.
URL https://doi.org/10.11909/j.issn.1671-5411.2014.04.008

[4] K. Häyrinen, K. Saranto, P. Nykänen, Definition, structure, content, use and impacts of electronic health records: a review of the research literature, International journal of medical informatics 77 (5) (2008) 291–304.
URL https://doi.org/10.1016/j.ijmedinf.2007.09.001

[5] E. Kim, S. M. Rubinstein, K. T. Nead, A. P. Wojcieszynski, P. E. Gabriel, J. L. Warner, The evolving use of electronic health records (ehr) for research, Seminars in radiation oncology 29 (4) (2019) 354–361.
URL https://doi.org/10.1016/j.semradonc.2019.05.010

[6] Y. Juhn, H. Liu, Artificial intelligence approaches using natural language processing to advance ehr-based clinical research, Journal of Allergy and Clinical Immunology 145 (2) (2020) 463–469.
URL https://doi.org/10.1016/j.jaci.2019.12.897

[7] E. Steinberg, K. Jung, J. A. Fries, C. K. Corbin, S. R. Pfohl, N. H. Shah, Language models are an effective representation learning technique for electronic health record data, Journal of biomedical informatics 113 (2021) 103637.
URL https://doi.org/10.1016/j.jbi.2020.103637

[8] M. Wornow, Y. Xu, R. Thapa, B. Patel, E. Steinberg, S. Fleming, M. A. Pfeffer, J. Fries, N. H. Shah, The shaky foundations of large language models and foundation models for electronic health records, npj digital medicine 6 (1) (2023) 135.
URL https://doi.org/10.1038/s41746-023-00879-8

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
URL https://dl.acm.org/doi/10.5555/3295222.3295349

[10] S. Nerella, S. Bandyopadhyay, J. Zhang, M. Contreras, S. Siegel, A. Bumin, B. Silva, J. Sena, B. Shickel, A. Bihorac, et al., Transformers and large language models in healthcare: A review, Artificial intelligence in medicine (2024) 102900.
URL https://doi.org/10.1016/j.artmed.2024.102900

[11] K. S. Kalyan, A. Rajasekharan, S. Sangeetha, Ammu: a survey of transformer-based biomedical pretrained

language models, Journal of biomedical informatics 126 (2022) 103982.
URL https://doi.org/10.1016/j.jbi.2021.103982

[12] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, H. Fu, Transformers in medical imaging: A survey, Medical image analysis 88 (2023) 102802.
URL https://doi.org/10.1016/j.media.2023.102802

[13] P. Nguyen, T. Tran, N. Wickramasinghe, S. Venkatesh, Deepr: a convolutional net for medical records, IEEE journal of biomedical and health informatics 21 (1) (2016) 22–30.
URL https://doi.org/10.1109/JBHI.2016.2633963

[14] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, Advances in neural information processing systems 29 (2016).
URL https://dl.acm.org/doi/10.5555/3157382.3157490

[15] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE transactions on neural networks 5 (2) (1994) 157–166.
URL https://doi.org/10.1109/72.279181

[16] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.
URL https://doi.org/10.1162/neco.1997.9.8.1735

[17] S. M. Al-Selwi, M. F. Hassan, S. J. Abdulkadir, A. Muneer, E. H. Sumiea, A. Alqushaibi, M. G. Ragab, Rnn-lstm: From applications to modeling techniques and beyond—systematic review, Journal of King Saud University-Computer and Information Sciences 36 (5) (2024) 102068.
URL https://doi.org/10.1016/j.jksuci.2024.102068

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
URL https://doi.org/10.18653/v1%2FN19-1423

[19] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, G. Salimi-Khorshidi, Behrt: transformer for electronic health records, Scientific reports 10 (1) (2020) 7155.
URL https://doi.org/10.1038/s41598-020-62922-y

[20] Y. Meng, W. Speier, M. K. Ong, C. W. Arnold, Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression, IEEE journal of biomedical and health informatics 25 (8) (2021) 3121–3129.
URL https://doi.org/10.1109/JBHI.2021.3063721

[21] C. Pang, X. Jiang, K. S. Kalluri, M. Spotnitz, R. Chen, A. Perotte, K. Natarajan, Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks, in: Machine Learning for Health, PMLR, 2021, pp. 239–260.
URL https://proceedings.mlr.press/v158/pang21a.html

[22] Z. Yang, A. Mitra, W. Liu, D. Berlowitz, H. Yu, Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records, Nature communications 14 (1) (2023) 7857.
URL https://doi.org/10.1038/s41467-023-43715-z

[23] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
URL https://doi.org/10.48550/arXiv.2302.13971

[24] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, in: COLM, 2023.
URL https://doi.org/10.48550/arXiv.2312.00752

[25] A. Fallahpour, M. Alinoori, W. Ye, X. Cao, A. Afkanpour, A. Krishnan, Ehrmamba: Towards generalizable and scalable foundation models for electronic health records, in: Proceedings of the 4th Machine Learning for Health Symposium, Vol. 259 of Proceedings of Machine Learning Research, PMLR, 2025, pp. 291–307.
URL https://proceedings.mlr.press/v259/fallahpour25a.html

[26] M. Wornow, S. Bedi, M. A. F. Hernandez, E. Steinberg, J. A. Fries, C. Ré, S. Koyejo, N. H. Shah, Context clues: Evaluating long context models for clinical prediction tasks on ehrs, in: ICLR, 2025.
URL https://openreview.net/forum?id=zg3ec1TdAP

[27] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, arXiv preprint arXiv:2001.08361 (2020).
URL https://doi.org/10.48550/arXiv.2001.08361

[28] H. Qu, L. Ning, R. An, W. Fan, T. Derr, H. Liu, X. Xu, Q. Li, A survey of mamba, arXiv preprint arXiv:2408.01129 (2024).
URL https://doi.org/10.48550/arXiv.2408.01129

[29] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, NPJ digital medicine 4 (1) (2021) 86.
URL https://doi.org/10.1038/s41746-021-00455-y

[30] M. Schaufelberger, S. Ekestubbe, S. Hultgren, H. Persson, A. Reimstad, M. Schaufelberger, A. Rosengren, Validity of heart failure diagnoses made in 2000–2012 in western sweden, ESC heart failure 7 (1) (2020) 37–46.
URL https://doi.org/10.1002/ehf2.12519

[31] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, G. T. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2025, pp. 2526–2547.
URL https://doi.org/10.18653/v1/2025.acl-long.127

[32] T. Dao, A. Gu, Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, in: ICML, 2024.
URL https://dl.acm.org/doi/10.5555/3692070.3692469

[33] M. Rupp, O. Peter, T. Pattipaka, Exbehrt: Extended transformer for electronic health records, in: International Workshop on Trustworthy Machine Learning for Healthcare, Springer, 2023, pp. 73–84.
URL https://doi.org/10.1007/978-3-031-39539-0_7

[34] E. Antikainen, J. Linnosmaa, A. Umer, N. Oksala, M. Eskola, M. van Gils, J. Hernesniemi, M. Gabbouj, Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records, Scientific Reports 13 (1) (2023) 3517.
URL https://doi.org/10.1038/s41598-023-30657-1

[35] M. Odgaard, K. V. Klein, S. M. Thysen, E. Jimenez-Solem, M. Sillesen, M. Nielsen, Core-behrt: A carefully optimized and rigorously evaluated behrt, in: Proceedings of the 9th Machine Learning for Healthcare Conference, Vol. 252 of Proceedings of Machine Learning Research, PMLR, 2024, pp. 1–33.
URL https://proceedings.mlr.press/v252/odgaard24a.html

[36] Y. Li, M. Mamouei, G. Salimi-Khorshidi, S. Rao, A. Hassaine, D. Canoy, T. Lukasiewicz, K. Rahimi, Hi-behrt: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records, IEEE journal of biomedical and health informatics 27 (2) (2022) 1106–1117.
URL https://doi.org/10.1109/JBHI.2022.3224727

[37] J. Shang, T. Ma, C. Xiao, J. Sun, Pre-training of graph augmented transformers for medication recommendation, in: International Joint Conference on Artificial Intelligence, 2019.
URL https://doi.org/10.24963/ijcai.2019%2F825

[38] S. Rao, M. Mamouei, G. Salimi-Khorshidi, Y. Li, R. Ramakrishnan, A. Hassaine, D. Canoy, K. Rahimi, Targeted-behrt: deep learning for observational causal inference on longitudinal electronic health records, IEEE Transactions on Neural Networks and Learning Systems 35 (4) (2022) 5027–5038.
URL https://doi.org/10.1109/tnnls.2022.3183864

[39] C. Pang, X. Jiang, N. P. Pavinkurve, K. S. Kalluri, E. L. Minto, J. Patterson, L. Zhang, G. Hripcsak, G. Gürsoy, N. Elhadad, et al., Cehr-gpt: Generating electronic health records with chronological patient timelines, arXiv preprint arXiv:2402.04400 (2024).
URL https://doi.org/10.48550/arXiv.2402.04400

[40] Z. Kraljevic, D. Bean, A. Shek, R. Bendayan, H. Hemingway, J. A. Yeung, A. Deng, A. Balston, J. Ross, E. 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) (2024) e281–e290.
URL https://doi.org/10.1016/S2589-7500(24)00025-6

[41] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).
URL https://dl.acm.org/doi/10.5555/3454287.3454804

[42] R. Waleffe, W. Byeon, D. Riach, B. Norick, V. Korthikanti, T. Dao, A. Gu, A. Hatamizadeh, S. Singh, D. Narayanan, et al., An empirical study of mamba-based language models, arXiv preprint arXiv:2406.07887 (2024).
URL https://doi.org/10.48550/arXiv.2406.07887

[43] X. Liu, C. Zhang, L. Zhang, Vision mamba: A comprehensive survey and taxonomy, arXiv preprint arXiv:2405.04404 (2024).

URL https://doi.org/10.48550/arXiv.2405.04404

[44] T. Darcet, M. Oquab, J. Mairal, P. Bojanowski, Vision transformers need registers, in: ICLR, 2024.
URL https://openreview.net/forum?id=2dnO3LLiJ1

[45] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
URL https://doi.org/10.1145/2939672.2939785

[46] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: ICLR, 2019.
URL https://openreview.net/forum?id=Bkg6RiCqY7

[47] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, Y. Liu, Roformer: Enhanced transformer with rotary position embedding, Neurocomputing 568 (2024) 127063.
URL https://doi.org/10.1016/j.neucom.2023.127063

[48] N. Shazeer, Glu variants improve transformer, arXiv preprint arXiv:2002.05202 (2020).
URL https://doi.org/10.48550/arXiv.2002.05202

[49] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, T. Liu, On layer normalization in the transformer architecture, in: ICML, 2020.
URL https://dl.acm.org/doi/10.5555/3524938.3525913

[50] B. Zhang, R. Sennrich, Root mean square layer normalization, Advances in neural information processing systems 32 (2019).
URL https://dl.acm.org/doi/abs/10.5555/3454287.3455397

Supplementary material

S1. Data

Table S1 highlights the extracted clinical concepts from EHRs. The eligibility criteria for clinical instability is shown in Table S2.

S2. Methods

Fig. S1 and Fig. S2 compare the affect on the number of tokens and longitudinal gap in days between first and last event for sequences varied by context length and time modifications respectively.

Table S1: The patient data consists of six different feature groups.

Concept	Subconcept
Demographics	Age, sex, body mass index
Diagnosis (DX)	Any ICD-10 code
Vital signs (VIT)	Body temperature, diastolic blood pressure, systolic blood pressure, pulse rate, respiratory rate, oxygen saturation
Laboratories (LAB)	Albumin, bilirubin, blood urea nitrogen, C-reactive protein, creatinine, ferritin, fasting glucose, plasma glucose, hemoglobin, glycated hemoglobin, N-terminal pro b-type natriuretic peptide, potassium, sodium, alanine transaminase, aspartate transaminase, troponin I, troponin T, uric acid
Procedures (PRO)	Any KVÅ code
Medications (MED)	Following ATC code levels: A02B, A08, B01, B03, C, H03, R03

Table S2: Eligibility criteria for unplanned hospitalizations in the clinical instability prediction task.

Hospitalization type	Eligibility criteria
Unplanned HF hospitalization	Service units: ÖS Med/Ger/Akut, Medicinklinik, Kardiologi, Medicin, Internmedicin, Specialistmedicin, MS Medicin o Akutmottagning, M210 Medicin Lidköping, Akutmedicin & Geriatri, M360 Kardiologi, Infektion-hematologi-hud, MS Geriatrik, Geriatrik neurologi och rehabilitering, Thorax, M220 Medicin Falköping, M420 AVA, Infektion, M330 Infektion, Njurmedicin, M350 Njurmedicin, Lungmedicin, Allergologi och Palliativ medicin, M130 Lungmedicin och neurologi, M120 Hematologi, M110 Stroke och rehabilitering, Akutmedicin, M230 Medicin Mariestad, K250 Palliativ vård, M160 Medicin slutenvård, M110 Stroke, neurologi och rehabilitering, M130 Lungmedicin och medicin, Klinik för nära vård, M365 Mag och tarm, M390 Geriatrik
Unplanned non-HF hospitalization	Surgical KVÅ codes: F, G, J, N, TF, V, XF, ZF; Medical KVÅ codes: AF, AG, AP, DF, DG, DP

S3. Additional results

This section presents additional experiments and results. Table S3 compares the performance of context lengths $C = 512$ and $C = 1024$ for the prediction of one-year mortality after latest hospitalization. For all conducted ablations the AUROC and Brier score are visualized in Figs. S3–S7 and numerical results of all metrics, including AUPRC, are presented in Tables S4– S18.

Table S3: Comparison of one-year mortality prediction after latest hospitalization for $C \in \{512, 1024\}$ and Medium-sized models. All metrics are reported with 95% CIs. Bold marks the **best AUPRC performance per context length and model** and bold underlined the **overall best AUPRC**.

Vocabulary	C	Tokens	Visits	ΔDays
$V_{b=10,i=3}$	512	369 (170, 512)	3 (2, 4)	356 (54, 1030)
$V_{b=10,i=3}$	1024	369 (170, 740)	3 (2, 5)	598 (103, 1383)

	AUPRC (↑)		AUROC (↑)		Brier (↓)	
Model	$C = 512$	$C = 1024$	$C = 512$	$C = 1024$	$C = 512$	$C = 1024$
XGBoost	**0.815** (0.802–0.829)	0.810 (0.795–0.826)	0.848 (0.839–0.857)	0.848 (0.839–0.857)	0.159 (0.154–0.164)	0.159 (0.154–0.164)
BERT	**0.835** (0.821–0.849)	0.834 (0.820–0.848)	0.866 (0.858–0.874)	0.865 (0.857–0.873)	0.150 (0.145–0.156)	0.152 (0.147–0.157)
XLNet	0.836 (0.822–0.849)	**0.842** (0.829–0.855)	0.863 (0.854–0.871)	0.869 (0.861–0.877)	0.151 (0.146–0.156)	0.153 (0.148–0.159)
MBERT	0.843 (0.830–0.856)	**0.845** (0.832–0.857)	0.871 (0.863–0.879)	0.871 (0.863–0.880)	0.147 (0.141–0.152)	0.148 (0.142–0.153)
Llama	0.845 (0.831–0.858)	**0.856** (0.843–0.868)	0.874 (0.866–0.882)	0.882 (0.875–0.890)	0.145 (0.140–0.151)	0.147 (0.142–0.152)
Mamba	0.851 (0.838–0.864)	**0.854** (0.841–0.866)	0.875 (0.868–0.883)	0.878 (0.870–0.886)	0.145 (0.140–0.150)	0.143 (0.138–0.148)
Mamba2	0.840 (0.826–0.855)	**0.860** (0.847–0.872)	0.872 (0.864–0.881)	0.884 (0.876–0.892)	0.146 (0.141–0.152)	0.138 (0.133–0.143)

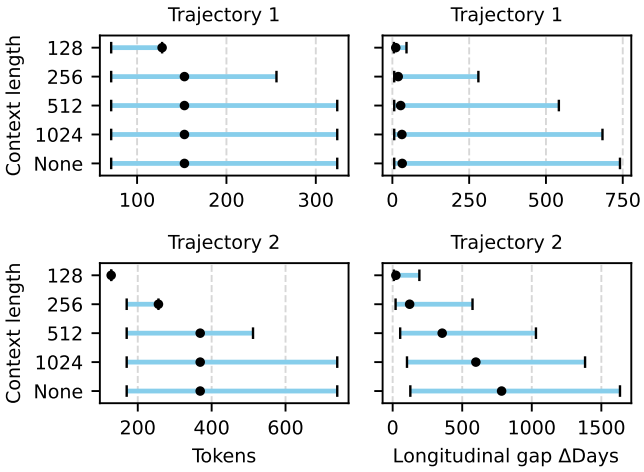


Fig. S1: Comparison of token and longitudinal gap properties visualized as the IQR (dot indicates median) determined by the selected context lengths or none. Trajectory 1 refers to the initial and trajectory 2 to the latest trajectory.

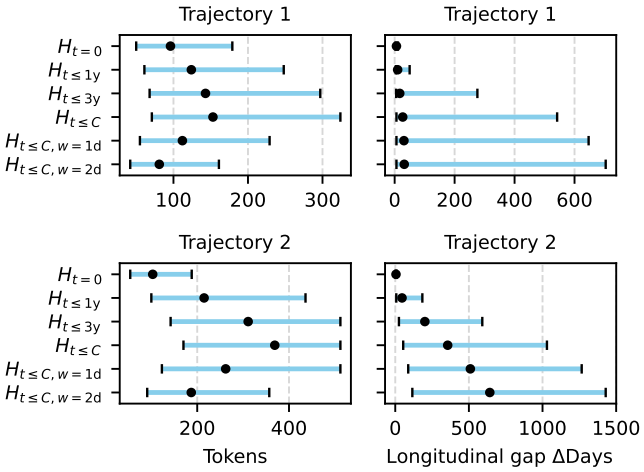
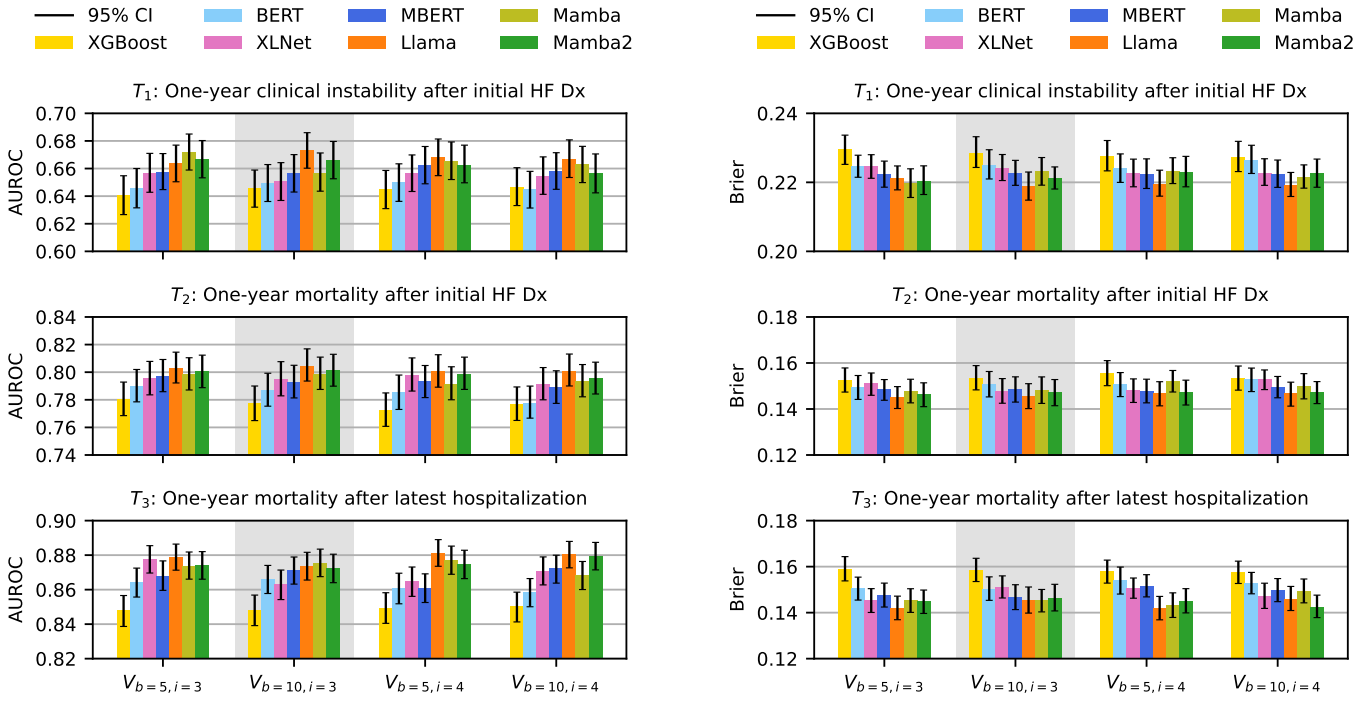
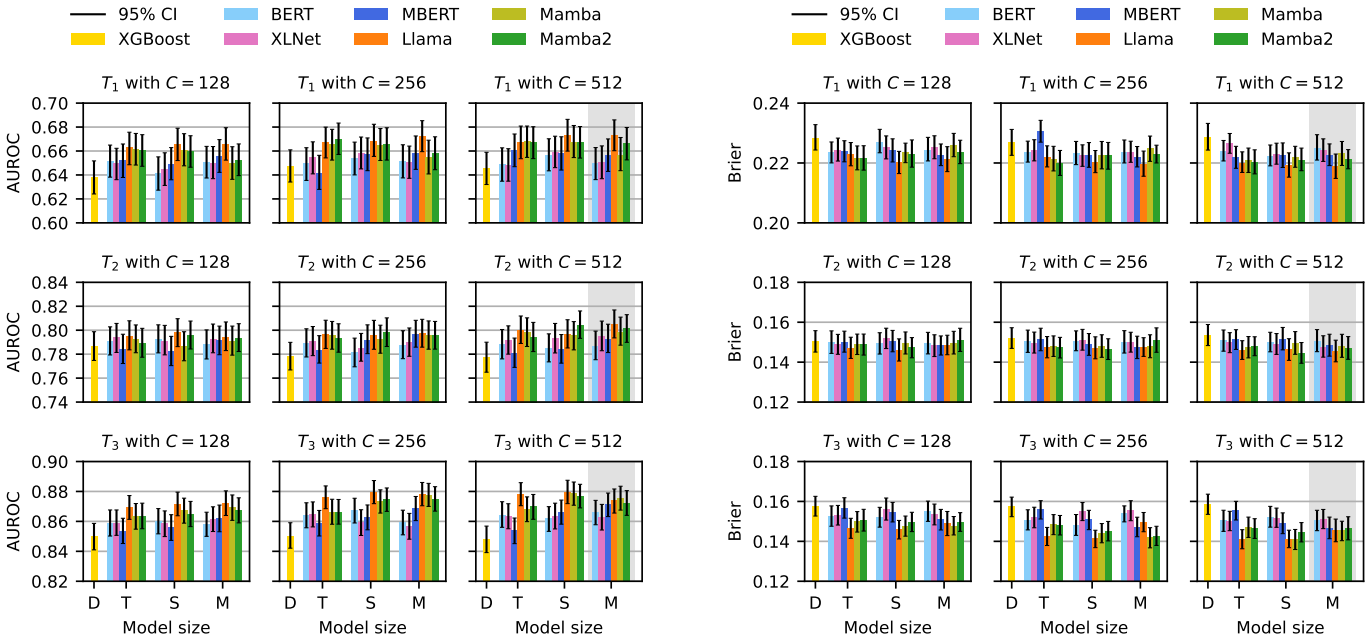


Fig. S2: Comparison of token and longitudinal gap properties visualized as the IQR (dot indicates median) determined by the selected time modification. Trajectory 1 refers to the initial and trajectory 2 to the latest trajectory.



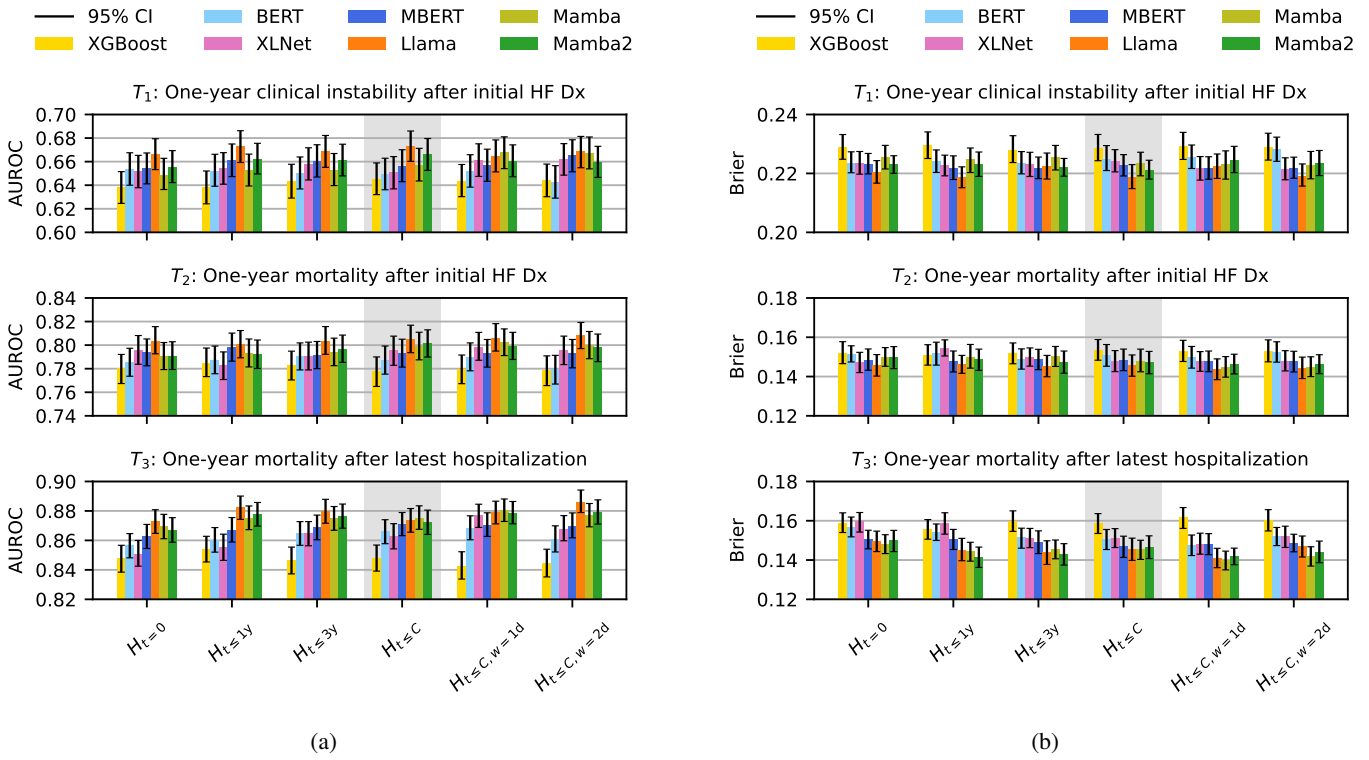
(a) (b)

Fig. S3: Ablation of the vocabulary V across three clinical tasks T evaluated by (a) AUROC (↑) and (b) Brier score (↓) for Medium-sized sequence models and $C = 512$. The resolution of the vocabulary increases from lowest (left) to highest (right). Gray background highlights common setup shared across all ablations.



(a) (b)

Fig. S4: Ablation of the context length C and model size evaluated by (a) AUROC (↑) and (b) Brier score (↓) across three different clinical tasks. Within each C the size of the sequence models is sorted by Tiny, Small, Medium and compared with the Default XGBoost configuration. Gray background highlights common setup shared across all ablations.

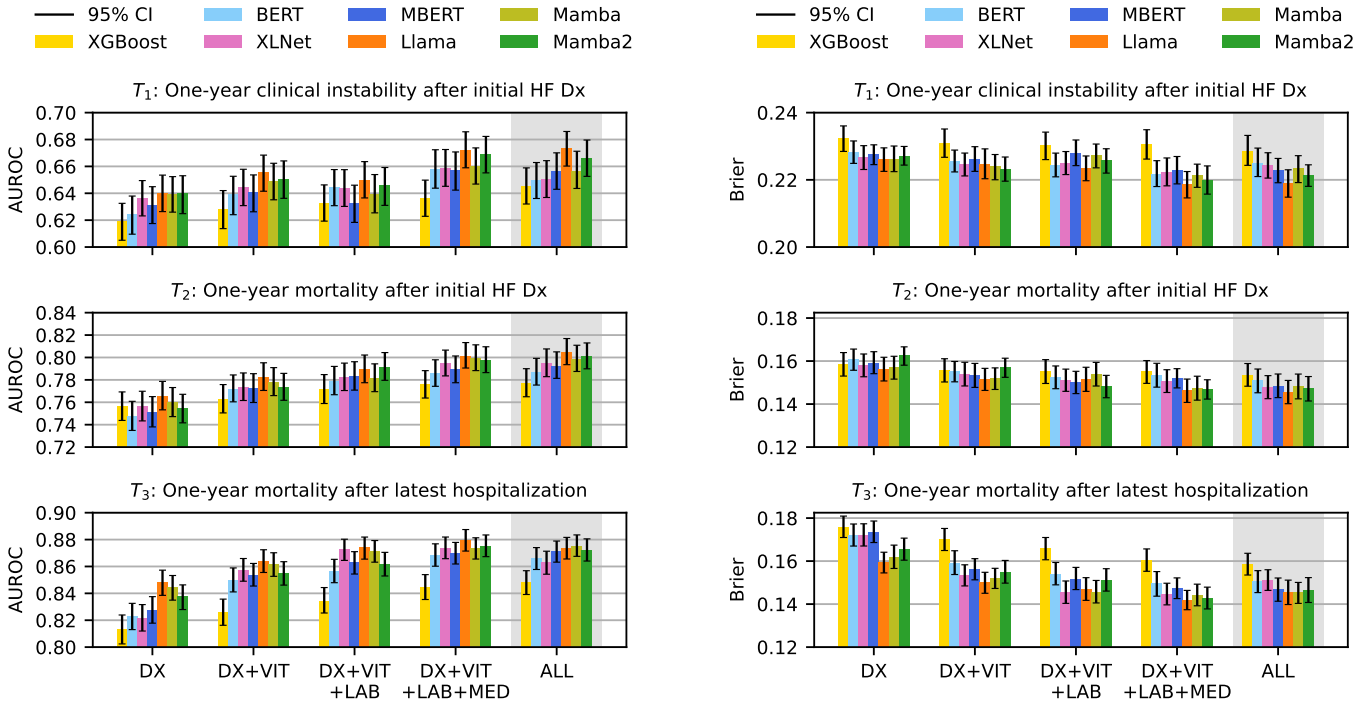


(a) (b)

Fig. S5: Ablation of the patient history through temporal modifications: The historical record is extended from the latest available admission (left) to a prolonged context (right). All modifications are evaluated by (a) AUROC (↑) and (b) Brier score (↓) using Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.

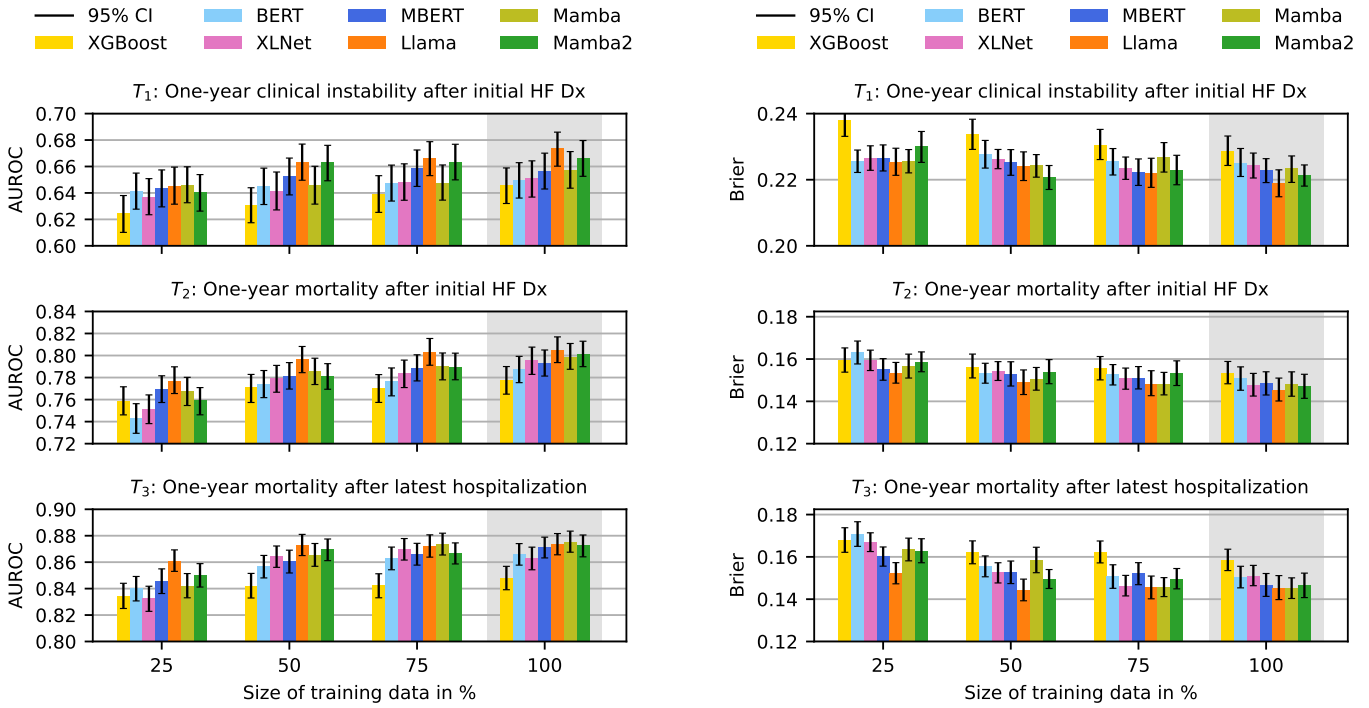
Table S4: Ablation of the vocabulary V across three clinical tasks T evaluated by AUPRC (↑) for Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.

Vocabulary	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx							
$V_{b=5,i=3}$	0.515 (0.494–0.535)	0.525 (0.506–0.547)	0.539 (0.518–0.558)	0.536 (0.516–0.558)	0.547 (0.527–0.567)	0.552 (0.532–0.573)	0.543 (0.522–0.564)
$V_{b=10,i=3}$	0.519 (0.497–0.540)	0.535 (0.516–0.556)	0.529 (0.508–0.549)	0.534 (0.513–0.556)	0.555 (0.535–0.575)	0.543 (0.523–0.565)	0.542 (0.521–0.562)
$V_{b=5,i=4}$	0.521 (0.500–0.541)	0.535 (0.515–0.555)	0.544 (0.524–0.564)	0.546 (0.526–0.567)	0.549 (0.529–0.570)	0.545 (0.526–0.566)	0.545 (0.525–0.567)
$V_{b=10,i=4}$	0.523 (0.502–0.544)	0.527 (0.507–0.547)	0.535 (0.515–0.555)	0.533 (0.514–0.553)	0.550 (0.529–0.570)	0.544 (0.523–0.564)	0.536 (0.516–0.556)
T_2: One-year mortality after initial HF Dx							
$V_{b=5,i=3}$	0.525 (0.501–0.552)	0.549 (0.525–0.574)	0.549 (0.523–0.575)	0.561 (0.537–0.586)	0.563 (0.537–0.589)	0.563 (0.537–0.588)	0.566 (0.542–0.591)
$V_{b=10,i=3}$	0.525 (0.500–0.550)	0.540 (0.516–0.564)	0.558 (0.533–0.584)	0.557 (0.531–0.583)	0.574 (0.550–0.599)	0.559 (0.534–0.585)	0.572 (0.547–0.596)
$V_{b=5,i=4}$	0.513 (0.489–0.538)	0.539 (0.514–0.565)	0.552 (0.526–0.579)	0.560 (0.535–0.584)	0.562 (0.536–0.586)	0.545 (0.521–0.570)	0.562 (0.536–0.586)
$V_{b=10,i=4}$	0.524 (0.498–0.549)	0.522 (0.496–0.545)	0.540 (0.514–0.566)	0.546 (0.521–0.570)	0.565 (0.541–0.592)	0.553 (0.528–0.579)	0.559 (0.533–0.585)
T_3: One-year clinical instability after latest hospitalization							
$V_{b=5,i=3}$	0.815 (0.801–0.829)	0.832 (0.818–0.845)	0.850 (0.838–0.863)	0.842 (0.828–0.854)	0.852 (0.839–0.864)	0.844 (0.831–0.857)	0.844 (0.831–0.857)
$V_{b=10,i=3}$	0.815 (0.802–0.829)	0.835 (0.821–0.849)	0.836 (0.822–0.849)	0.843 (0.830–0.856)	0.845 (0.831–0.858)	0.851 (0.838–0.864)	0.840 (0.826–0.855)
$V_{b=5,i=4}$	0.815 (0.802–0.830)	0.827 (0.813–0.841)	0.833 (0.819–0.847)	0.826 (0.812–0.840)	0.854 (0.843–0.867)	0.850 (0.837–0.862)	0.846 (0.833–0.858)
$V_{b=10,i=4}$	0.816 (0.802–0.831)	0.828 (0.814–0.841)	0.843 (0.831–0.856)	0.844 (0.832–0.857)	0.853 (0.841–0.866)	0.841 (0.828–0.853)	0.853 (0.840–0.865)



(a) (b)

Fig. S6: Augmentation of the DX group through the incremental inclusion of the respective clinical, continuing until all available concepts are included. Each augmentation is evaluated by (a) AUROC (↑) and (b) Brier score (↓) using Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.



(a) (b)

Fig. S7: Ablation of different training sizes for Medium-sized sequence models and $C = 512$ evaluated by (a) AUROC (↑) and (b) Brier score (↓). Gray background highlights common setup shared across all ablations.

Table S5: Ablation of the vocabulary V across three clinical tasks T evaluated by AUROC (↑) for Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.

Vocabulary	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx							
$V_{b=5,i=3}$	0.641 (0.627–0.655)	0.646 (0.632–0.660)	0.657 (0.643–0.671)	0.658 (0.645–0.671)	0.664 (0.650–0.677)	0.672 (0.659–0.685)	0.667 (0.653–0.680)
$V_{b=10,i=3}$	0.646 (0.632–0.659)	0.650 (0.636–0.663)	0.651 (0.637–0.664)	0.656 (0.643–0.670)	0.673 (0.660–0.686)	0.657 (0.644–0.671)	0.666 (0.653–0.680)
$V_{b=5,i=4}$	0.645 (0.631–0.659)	0.650 (0.636–0.663)	0.657 (0.643–0.670)	0.662 (0.649–0.676)	0.669 (0.655–0.681)	0.666 (0.652–0.679)	0.663 (0.650–0.677)
$V_{b=10,i=4}$	0.647 (0.633–0.661)	0.645 (0.631–0.658)	0.655 (0.641–0.668)	0.658 (0.645–0.671)	0.667 (0.654–0.681)	0.663 (0.650–0.676)	0.657 (0.642–0.671)
T_2: One-year mortality after initial HF Dx							
$V_{b=5,i=3}$	0.781 (0.769–0.793)	0.790 (0.779–0.802)	0.795 (0.784–0.808)	0.797 (0.786–0.809)	0.803 (0.792–0.815)	0.798 (0.787–0.810)	0.801 (0.789–0.812)
$V_{b=10,i=3}$	0.778 (0.765–0.790)	0.787 (0.775–0.799)	0.795 (0.783–0.808)	0.793 (0.781–0.805)	0.805 (0.794–0.817)	0.799 (0.787–0.811)	0.801 (0.790–0.813)
$V_{b=5,i=4}$	0.772 (0.761–0.785)	0.785 (0.773–0.798)	0.798 (0.786–0.810)	0.793 (0.782–0.805)	0.801 (0.789–0.813)	0.791 (0.780–0.804)	0.799 (0.788–0.811)
$V_{b=10,i=4}$	0.777 (0.765–0.789)	0.778 (0.767–0.790)	0.791 (0.780–0.803)	0.789 (0.777–0.801)	0.801 (0.790–0.813)	0.793 (0.782–0.806)	0.795 (0.784–0.807)
T_3: One-year clinical instability after latest hospitalization							
$V_{b=5,i=3}$	0.848 (0.839–0.857)	0.864 (0.856–0.872)	0.878 (0.870–0.885)	0.868 (0.860–0.877)	0.879 (0.871–0.886)	0.874 (0.866–0.882)	0.874 (0.866–0.882)
$V_{b=10,i=3}$	0.848 (0.839–0.857)	0.866 (0.858–0.874)	0.863 (0.854–0.871)	0.871 (0.863–0.879)	0.874 (0.866–0.882)	0.875 (0.868–0.883)	0.872 (0.864–0.881)
$V_{b=5,i=4}$	0.849 (0.840–0.858)	0.861 (0.852–0.870)	0.865 (0.856–0.873)	0.861 (0.853–0.869)	0.881 (0.874–0.889)	0.877 (0.869–0.885)	0.875 (0.866–0.883)
$V_{b=10,i=4}$	0.850 (0.841–0.859)	0.858 (0.850–0.866)	0.871 (0.863–0.879)	0.872 (0.864–0.880)	0.880 (0.873–0.888)	0.868 (0.860–0.876)	0.880 (0.871–0.887)

Table S6: Ablation of the vocabulary V across three clinical tasks T evaluated by Brier score (↓) for Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.

Vocabulary	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx							
$V_{b=5,i=3}$	0.230 (0.225–0.234)	0.225 (0.221–0.228)	0.225 (0.221–0.228)	0.222 (0.219–0.226)	0.221 (0.218–0.225)	0.220 (0.216–0.224)	0.221 (0.216–0.225)
$V_{b=10,i=3}$	0.229 (0.224–0.233)	0.225 (0.221–0.229)	0.224 (0.221–0.228)	0.223 (0.219–0.226)	0.219 (0.215–0.223)	0.223 (0.219–0.227)	0.221 (0.218–0.224)
$V_{b=5,i=4}$	0.228 (0.223–0.232)	0.224 (0.220–0.228)	0.223 (0.219–0.227)	0.223 (0.218–0.227)	0.220 (0.216–0.224)	0.223 (0.220–0.227)	0.223 (0.219–0.228)
$V_{b=10,i=4}$	0.227 (0.223–0.232)	0.227 (0.223–0.231)	0.223 (0.219–0.227)	0.223 (0.219–0.226)	0.219 (0.216–0.223)	0.222 (0.218–0.225)	0.223 (0.219–0.227)
T_2: One-year mortality after initial HF Dx							
$V_{b=5,i=3}$	0.153 (0.147–0.158)	0.149 (0.144–0.155)	0.151 (0.146–0.156)	0.148 (0.144–0.153)	0.145 (0.140–0.150)	0.148 (0.143–0.153)	0.146 (0.141–0.151)
$V_{b=10,i=3}$	0.153 (0.148–0.159)	0.151 (0.145–0.156)	0.148 (0.142–0.153)	0.148 (0.143–0.154)	0.146 (0.140–0.151)	0.148 (0.142–0.154)	0.147 (0.141–0.153)
$V_{b=5,i=4}$	0.155 (0.150–0.161)	0.151 (0.145–0.156)	0.148 (0.143–0.153)	0.148 (0.143–0.153)	0.147 (0.141–0.152)	0.152 (0.147–0.157)	0.147 (0.142–0.153)
$V_{b=10,i=4}$	0.153 (0.148–0.159)	0.153 (0.148–0.158)	0.153 (0.148–0.157)	0.149 (0.145–0.154)	0.147 (0.141–0.152)	0.150 (0.144–0.155)	0.147 (0.142–0.152)
T_3: One-year clinical instability after latest hospitalization							
$V_{b=5,i=3}$	0.159 (0.154–0.164)	0.151 (0.145–0.155)	0.145 (0.140–0.150)	0.148 (0.142–0.153)	0.142 (0.137–0.147)	0.145 (0.140–0.150)	0.145 (0.140–0.150)
$V_{b=10,i=3}$	0.159 (0.154–0.164)	0.150 (0.145–0.156)	0.151 (0.146–0.156)	0.147 (0.141–0.152)	0.145 (0.140–0.151)	0.145 (0.140–0.150)	0.146 (0.141–0.152)
$V_{b=5,i=4}$	0.158 (0.153–0.163)	0.154 (0.148–0.160)	0.151 (0.146–0.155)	0.152 (0.147–0.157)	0.142 (0.137–0.147)	0.143 (0.138–0.149)	0.145 (0.140–0.150)
$V_{b=10,i=4}$	0.157 (0.153–0.162)	0.153 (0.148–0.157)	0.147 (0.142–0.153)	0.150 (0.145–0.155)	0.146 (0.141–0.151)	0.149 (0.144–0.155)	0.143 (0.138–0.148)

Table S7: Ablation of the context length C and model size S across three clinical tasks T evaluated by AUPRC (↑). Here, T, S, M denote the Tiny, Small, Medium model sizes. XGBoost's default configuration is merged with size T for readability. Gray background highlights common setup shared across all ablations.

C	S	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx								
128	T	0.517 (0.496–0.538)	0.526 (0.505–0.547)	0.529 (0.508–0.550)	0.530 (0.509–0.551)	0.542 (0.522–0.563)	0.538 (0.517–0.559)	0.539 (0.518–0.560)
128	S	-	0.511 (0.491–0.533)	0.525 (0.504–0.548)	0.528 (0.507–0.549)	0.543 (0.523–0.563)	0.539 (0.518–0.560)	0.538 (0.517–0.561)
128	M	-	0.533 (0.511–0.555)	0.529 (0.507–0.551)	0.536 (0.515–0.558)	0.543 (0.523–0.563)	0.529 (0.508–0.550)	0.531 (0.510–0.551)
256	T	0.522 (0.501–0.542)	0.531 (0.512–0.551)	0.534 (0.513–0.556)	0.520 (0.500–0.542)	0.551 (0.530–0.572)	0.542 (0.522–0.562)	0.545 (0.526–0.567)
256	S	-	0.536 (0.515–0.556)	0.532 (0.512–0.553)	0.539 (0.519–0.561)	0.551 (0.531–0.572)	0.543 (0.523–0.564)	0.543 (0.522–0.564)
256	M	-	0.527 (0.507–0.547)	0.530 (0.508–0.551)	0.540 (0.520–0.562)	0.554 (0.534–0.575)	0.537 (0.516–0.559)	0.537 (0.516–0.558)
512	T	0.519 (0.497–0.540)	0.526 (0.505–0.545)	0.535 (0.513–0.556)	0.541 (0.520–0.562)	0.550 (0.529–0.570)	0.546 (0.526–0.566)	0.549 (0.529–0.570)
512	S	-	0.539 (0.519–0.560)	0.537 (0.516–0.559)	0.538 (0.519–0.559)	0.553 (0.533–0.574)	0.545 (0.525–0.566)	0.547 (0.526–0.567)
512	M	-	0.535 (0.516–0.556)	0.529 (0.508–0.549)	0.534 (0.513–0.556)	0.555 (0.535–0.575)	0.543 (0.523–0.565)	0.542 (0.521–0.562)
T_2: One-year mortality after initial HF Dx								
128	T	0.538 (0.512–0.563)	0.546 (0.518–0.572)	0.553 (0.526–0.579)	0.544 (0.517–0.569)	0.564 (0.538–0.589)	0.557 (0.531–0.583)	0.553 (0.525–0.580)
128	S	-	0.555 (0.530–0.580)	0.555 (0.529–0.582)	0.542 (0.516–0.567)	0.569 (0.543–0.594)	0.547 (0.522–0.574)	0.560 (0.535–0.585)
128	M	-	0.549 (0.523–0.575)	0.556 (0.530–0.581)	0.562 (0.536–0.588)	0.556 (0.529–0.582)	0.548 (0.521–0.574)	0.560 (0.535–0.585)
256	T	0.539 (0.514–0.562)	0.549 (0.523–0.575)	0.554 (0.527–0.578)	0.538 (0.513–0.564)	0.565 (0.539–0.589)	0.561 (0.535–0.586)	0.562 (0.536–0.587)
256	S	-	0.539 (0.513–0.564)	0.538 (0.513–0.563)	0.557 (0.531–0.582)	0.568 (0.542–0.592)	0.563 (0.538–0.588)	0.561 (0.534–0.587)
256	M	-	0.542 (0.516–0.567)	0.553 (0.528–0.579)	0.565 (0.540–0.591)	0.558 (0.532–0.582)	0.559 (0.532–0.585)	0.569 (0.544–0.593)
512	T	0.525 (0.500–0.550)	0.538 (0.512–0.564)	0.550 (0.524–0.575)	0.532 (0.507–0.560)	0.566 (0.540–0.592)	0.561 (0.534–0.587)	0.558 (0.533–0.584)
512	S	-	0.540 (0.513–0.565)	0.554 (0.528–0.580)	0.540 (0.516–0.565)	0.571 (0.545–0.595)	0.558 (0.532–0.583)	0.571 (0.544–0.597)
512	M	-	0.540 (0.516–0.564)	0.558 (0.533–0.584)	0.557 (0.531–0.583)	0.574 (0.550–0.599)	0.559 (0.534–0.585)	0.572 (0.547–0.596)
T_3: One-year clinical instability after latest hospitalization								
128	T	0.816 (0.801–0.829)	0.829 (0.814–0.842)	0.828 (0.813–0.842)	0.823 (0.808–0.836)	0.841 (0.828–0.854)	0.835 (0.821–0.848)	0.835 (0.821–0.849)
128	S	-	0.829 (0.816–0.843)	0.832 (0.818–0.845)	0.824 (0.810–0.838)	0.845 (0.832–0.857)	0.839 (0.825–0.852)	0.832 (0.817–0.845)
128	M	-	0.826 (0.812–0.840)	0.834 (0.822–0.847)	0.834 (0.821–0.847)	0.847 (0.834–0.859)	0.841 (0.827–0.854)	0.838 (0.824–0.851)
256	T	0.818 (0.803–0.833)	0.832 (0.818–0.846)	0.832 (0.818–0.846)	0.829 (0.815–0.842)	0.849 (0.835–0.861)	0.840 (0.826–0.853)	0.836 (0.823–0.850)
256	S	-	0.840 (0.827–0.852)	0.827 (0.814–0.842)	0.833 (0.819–0.846)	0.851 (0.838–0.863)	0.846 (0.833–0.859)	0.844 (0.830–0.857)
256	M	-	0.828 (0.815–0.842)	0.827 (0.813–0.840)	0.843 (0.831–0.856)	0.850 (0.837–0.862)	0.852 (0.839–0.865)	0.846 (0.832–0.860)
512	T	0.815 (0.802–0.829)	0.832 (0.818–0.847)	0.833 (0.819–0.846)	0.816 (0.800–0.830)	0.849 (0.836–0.862)	0.840 (0.826–0.853)	0.840 (0.827–0.854)
512	S	-	0.831 (0.817–0.845)	0.834 (0.821–0.847)	0.839 (0.826–0.852)	0.853 (0.840–0.865)	0.853 (0.841–0.865)	0.848 (0.835–0.861)
512	M	-	0.835 (0.821–0.849)	0.836 (0.822–0.849)	0.843 (0.830–0.856)	0.845 (0.831–0.858)	0.851 (0.838–0.864)	0.840 (0.826–0.855)

Table S8: Ablation of the context length C and model size S across three clinical tasks T evaluated by AUROC (↑). Here, T, S, M denote the Tiny, Small, Medium model sizes. XGBoost's default configuration is merged with size T for readability. Gray background highlights common setup shared across all ablations.

C	S	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx								
128	T	0.638 (0.624–0.652)	0.652 (0.638–0.665)	0.649 (0.636–0.662)	0.652 (0.638–0.666)	0.663 (0.649–0.676)	0.662 (0.648–0.675)	0.661 (0.647–0.674)
128	S	-	0.641 (0.627–0.655)	0.645 (0.631–0.659)	0.649 (0.636–0.663)	0.666 (0.652–0.679)	0.661 (0.648–0.674)	0.660 (0.646–0.673)
128	M	-	0.651 (0.637–0.664)	0.650 (0.637–0.664)	0.656 (0.642–0.670)	0.666 (0.652–0.679)	0.650 (0.636–0.664)	0.652 (0.639–0.666)
256	T	0.648 (0.634–0.661)	0.650 (0.635–0.663)	0.655 (0.641–0.668)	0.642 (0.628–0.656)	0.668 (0.654–0.680)	0.666 (0.652–0.678)	0.670 (0.657–0.683)
256	S	-	0.654 (0.640–0.667)	0.658 (0.645–0.672)	0.657 (0.644–0.671)	0.669 (0.656–0.682)	0.665 (0.652–0.679)	0.666 (0.652–0.679)
256	M	-	0.651 (0.638–0.665)	0.651 (0.637–0.664)	0.658 (0.645–0.673)	0.673 (0.659–0.685)	0.655 (0.641–0.669)	0.658 (0.644–0.672)
512	T	0.646 (0.632–0.659)	0.649 (0.635–0.663)	0.648 (0.635–0.662)	0.661 (0.647–0.674)	0.668 (0.654–0.681)	0.668 (0.655–0.681)	0.667 (0.654–0.680)
512	S	-	0.657 (0.643–0.670)	0.659 (0.646–0.672)	0.658 (0.644–0.672)	0.673 (0.660–0.686)	0.668 (0.655–0.681)	0.667 (0.654–0.680)
512	M	-	0.650 (0.636–0.663)	0.651 (0.637–0.664)	0.656 (0.643–0.670)	0.673 (0.660–0.686)	0.657 (0.644–0.671)	0.666 (0.653–0.680)
T_2: One-year mortality after initial HF Dx								
128	T	0.786 (0.775–0.799)	0.791 (0.779–0.803)	0.794 (0.781–0.806)	0.784 (0.772–0.797)	0.795 (0.784–0.808)	0.792 (0.781–0.804)	0.789 (0.777–0.802)
128	S	-	0.792 (0.781–0.804)	0.791 (0.779–0.804)	0.783 (0.770–0.795)	0.798 (0.786–0.810)	0.786 (0.774–0.799)	0.796 (0.784–0.808)
128	M	-	0.788 (0.776–0.800)	0.793 (0.781–0.805)	0.791 (0.780–0.803)	0.794 (0.783–0.807)	0.791 (0.779–0.803)	0.794 (0.782–0.805)
256	T	0.779 (0.767–0.790)	0.789 (0.778–0.801)	0.791 (0.779–0.803)	0.783 (0.773–0.795)	0.796 (0.785–0.808)	0.796 (0.784–0.807)	0.793 (0.782–0.805)
256	S	-	0.781 (0.769–0.793)	0.785 (0.774–0.797)	0.792 (0.780–0.804)	0.796 (0.784–0.808)	0.792 (0.781–0.804)	0.798 (0.786–0.810)
256	M	-	0.788 (0.776–0.800)	0.790 (0.778–0.802)	0.797 (0.786–0.808)	0.797 (0.786–0.809)	0.796 (0.784–0.808)	0.796 (0.784–0.807)
512	T	0.778 (0.765–0.790)	0.788 (0.776–0.801)	0.792 (0.780–0.804)	0.781 (0.769–0.793)	0.800 (0.789–0.812)	0.798 (0.786–0.810)	0.794 (0.782–0.807)
512	S	-	0.785 (0.774–0.797)	0.793 (0.781–0.806)	0.784 (0.772–0.796)	0.797 (0.785–0.809)	0.796 (0.784–0.807)	0.804 (0.793–0.816)
512	M	-	0.787 (0.775–0.799)	0.795 (0.783–0.808)	0.793 (0.781–0.805)	0.805 (0.794–0.817)	0.799 (0.787–0.811)	0.801 (0.790–0.813)
T_3: One-year clinical instability after latest hospitalization								
128	T	0.850 (0.841–0.859)	0.859 (0.850–0.868)	0.859 (0.851–0.868)	0.854 (0.845–0.862)	0.869 (0.861–0.877)	0.863 (0.855–0.872)	0.864 (0.855–0.872)
128	S	-	0.860 (0.852–0.869)	0.858 (0.850–0.867)	0.856 (0.847–0.864)	0.871 (0.864–0.880)	0.867 (0.859–0.876)	0.865 (0.857–0.873)
128	M	-	0.858 (0.850–0.866)	0.862 (0.853–0.870)	0.862 (0.853–0.871)	0.872 (0.864–0.880)	0.869 (0.861–0.878)	0.867 (0.859–0.876)
256	T	0.850 (0.842–0.859)	0.864 (0.856–0.872)	0.865 (0.856–0.873)	0.859 (0.850–0.867)	0.876 (0.868–0.884)	0.866 (0.858–0.875)	0.866 (0.858–0.875)
256	S	-	0.867 (0.859–0.875)	0.859 (0.851–0.868)	0.863 (0.854–0.871)	0.879 (0.872–0.887)	0.874 (0.866–0.881)	0.875 (0.867–0.882)
256	M	-	0.859 (0.851–0.867)	0.857 (0.848–0.865)	0.869 (0.861–0.877)	0.878 (0.870–0.886)	0.878 (0.870–0.885)	0.875 (0.867–0.883)
512	T	0.848 (0.839–0.857)	0.864 (0.856–0.873)	0.863 (0.855–0.872)	0.854 (0.845–0.862)	0.878 (0.870–0.886)	0.868 (0.860–0.876)	0.870 (0.861–0.878)
512	S	-	0.862 (0.853–0.870)	0.863 (0.855–0.872)	0.866 (0.858–0.874)	0.880 (0.872–0.887)	0.879 (0.871–0.886)	0.877 (0.869–0.885)
512	M	-	0.866 (0.858–0.874)	0.863 (0.854–0.871)	0.871 (0.863–0.879)	0.874 (0.866–0.882)	0.875 (0.868–0.883)	0.872 (0.864–0.881)

Table S9: Ablation of the context length C and model size S across three clinical tasks T evaluated by Brier score (↓). Here, T, S, M denote the Tiny, Small, Medium model sizes. XGBoost's default configuration is merged with size T for readability. Gray background highlights common setup shared across all ablations.

C	S	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx								
128	T	0.228 (0.224–0.233)	0.223 (0.220–0.227)	0.224 (0.221–0.228)	0.224 (0.220–0.227)	0.223 (0.219–0.227)	0.222 (0.218–0.226)	0.222 (0.218–0.226)
128	S	-	0.227 (0.223–0.231)	0.225 (0.221–0.229)	0.224 (0.220–0.228)	0.220 (0.216–0.224)	0.223 (0.220–0.227)	0.223 (0.219–0.228)
128	M	-	0.224 (0.221–0.229)	0.225 (0.221–0.229)	0.223 (0.219–0.226)	0.221 (0.217–0.225)	0.226 (0.222–0.230)	0.224 (0.220–0.228)
256	T	0.227 (0.223–0.231)	0.224 (0.220–0.227)	0.224 (0.221–0.228)	0.231 (0.227–0.234)	0.222 (0.219–0.226)	0.221 (0.217–0.226)	0.220 (0.216–0.224)
256	S	-	0.223 (0.219–0.227)	0.222 (0.219–0.226)	0.223 (0.219–0.227)	0.220 (0.217–0.224)	0.223 (0.218–0.227)	0.222 (0.218–0.227)
256	M	-	0.224 (0.220–0.228)	0.224 (0.220–0.227)	0.222 (0.219–0.226)	0.220 (0.216–0.224)	0.225 (0.221–0.229)	0.223 (0.220–0.226)
512	T	0.229 (0.224–0.233)	0.224 (0.221–0.228)	0.227 (0.223–0.230)	0.222 (0.218–0.226)	0.220 (0.217–0.224)	0.221 (0.217–0.225)	0.220 (0.216–0.224)
512	S	-	0.222 (0.219–0.226)	0.223 (0.220–0.227)	0.222 (0.219–0.227)	0.219 (0.215–0.223)	0.222 (0.219–0.226)	0.221 (0.217–0.225)
512	M	-	0.225 (0.221–0.229)	0.224 (0.221–0.228)	0.223 (0.219–0.226)	0.219 (0.215–0.223)	0.223 (0.219–0.227)	0.221 (0.218–0.224)
T_2: One-year mortality after initial HF Dx								
128	T	0.151 (0.145–0.156)	0.150 (0.144–0.156)	0.149 (0.144–0.154)	0.150 (0.145–0.156)	0.147 (0.142–0.152)	0.149 (0.144–0.154)	0.149 (0.144–0.154)
128	S	-	0.150 (0.144–0.155)	0.152 (0.147–0.157)	0.151 (0.145–0.156)	0.146 (0.141–0.152)	0.149 (0.144–0.155)	0.147 (0.142–0.152)
128	M	-	0.150 (0.144–0.155)	0.148 (0.143–0.154)	0.148 (0.144–0.153)	0.149 (0.144–0.153)	0.150 (0.144–0.155)	0.151 (0.145–0.157)
256	T	0.152 (0.147–0.157)	0.151 (0.145–0.156)	0.150 (0.145–0.155)	0.151 (0.146–0.157)	0.147 (0.142–0.153)	0.148 (0.143–0.153)	0.147 (0.142–0.152)
256	S	-	0.151 (0.146–0.156)	0.151 (0.146–0.156)	0.149 (0.143–0.154)	0.147 (0.142–0.153)	0.148 (0.142–0.154)	0.147 (0.141–0.152)
256	M	-	0.150 (0.145–0.156)	0.150 (0.145–0.155)	0.148 (0.142–0.153)	0.148 (0.143–0.152)	0.148 (0.142–0.154)	0.151 (0.145–0.157)
512	T	0.153 (0.148–0.159)	0.151 (0.145–0.156)	0.150 (0.145–0.155)	0.151 (0.146–0.156)	0.146 (0.141–0.151)	0.148 (0.142–0.153)	0.148 (0.143–0.153)
512	S	-	0.150 (0.145–0.155)	0.149 (0.144–0.154)	0.152 (0.146–0.157)	0.146 (0.141–0.152)	0.149 (0.144–0.155)	0.145 (0.139–0.150)
512	M	-	0.151 (0.145–0.156)	0.148 (0.142–0.153)	0.148 (0.143–0.154)	0.146 (0.140–0.151)	0.148 (0.142–0.154)	0.147 (0.141–0.153)
T_3: One-year clinical instability after latest hospitalization								
128	T	0.158 (0.153–0.163)	0.153 (0.148–0.158)	0.153 (0.148–0.158)	0.157 (0.152–0.162)	0.147 (0.142–0.152)	0.150 (0.145–0.155)	0.151 (0.145–0.156)
128	S	-	0.152 (0.147–0.157)	0.156 (0.151–0.162)	0.155 (0.150–0.160)	0.146 (0.141–0.150)	0.148 (0.143–0.153)	0.150 (0.145–0.155)
128	M	-	0.155 (0.150–0.160)	0.154 (0.148–0.159)	0.151 (0.146–0.156)	0.149 (0.143–0.155)	0.148 (0.143–0.152)	0.150 (0.145–0.154)
256	T	0.157 (0.152–0.162)	0.151 (0.146–0.156)	0.152 (0.147–0.157)	0.156 (0.151–0.160)	0.142 (0.138–0.147)	0.148 (0.143–0.153)	0.148 (0.143–0.153)
256	S	-	0.148 (0.143–0.153)	0.155 (0.150–0.159)	0.151 (0.146–0.156)	0.141 (0.137–0.146)	0.144 (0.139–0.149)	0.145 (0.140–0.150)
256	M	-	0.154 (0.150–0.158)	0.155 (0.150–0.160)	0.147 (0.142–0.152)	0.150 (0.145–0.154)	0.142 (0.137–0.147)	0.143 (0.138–0.148)
512	T	0.159 (0.154–0.164)	0.151 (0.145–0.156)	0.150 (0.145–0.155)	0.155 (0.151–0.160)	0.141 (0.136–0.146)	0.147 (0.142–0.152)	0.146 (0.141–0.151)
512	S	-	0.152 (0.147–0.158)	0.152 (0.147–0.157)	0.149 (0.144–0.154)	0.141 (0.137–0.146)	0.141 (0.136–0.146)	0.144 (0.140–0.149)
512	M	-	0.150 (0.145–0.156)	0.151 (0.146–0.156)	0.147 (0.141–0.152)	0.145 (0.140–0.151)	0.145 (0.140–0.150)	0.146 (0.141–0.152)

Table S10: Ablation of the history H across three clinical tasks T evaluated by AUPRC (↑) for Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.

History	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx							
$H_{t=0}$	0.515 (0.495–0.536)	0.531 (0.510–0.552)	0.531 (0.511–0.552)	0.536 (0.516–0.557)	0.541 (0.521–0.562)	0.526 (0.506–0.546)	0.528 (0.508–0.549)
$H_{t\leq 1y}$	0.518 (0.498–0.539)	0.535 (0.515–0.555)	0.535 (0.514–0.555)	0.542 (0.521–0.561)	0.557 (0.537–0.576)	0.533 (0.513–0.553)	0.536 (0.516–0.557)
$H_{t\leq 3y}$	0.523 (0.502–0.545)	0.536 (0.515–0.556)	0.541 (0.519–0.561)	0.543 (0.522–0.563)	0.555 (0.535–0.576)	0.534 (0.513–0.554)	0.541 (0.520–0.560)
$H_{t\leq C}$	0.519 (0.497–0.540)	0.535 (0.516–0.556)	0.529 (0.508–0.549)	0.534 (0.513–0.556)	0.555 (0.535–0.575)	0.543 (0.523–0.565)	0.542 (0.521–0.562)
$H_{t\leq C,w=1d}$	0.514 (0.494–0.536)	0.536 (0.515–0.557)	0.543 (0.522–0.563)	0.534 (0.514–0.554)	0.549 (0.529–0.570)	0.543 (0.523–0.564)	0.533 (0.513–0.553)
$H_{t\leq C,w=2d}$	0.515 (0.495–0.537)	0.518 (0.496–0.538)	0.543 (0.523–0.563)	0.543 (0.521–0.563)	0.551 (0.530–0.572)	0.546 (0.526–0.566)	0.536 (0.516–0.556)
T_2: One-year mortality after initial HF Dx							
$H_{t=0}$	0.533 (0.506–0.559)	0.542 (0.518–0.567)	0.558 (0.533–0.583)	0.562 (0.538–0.587)	0.575 (0.551–0.600)	0.545 (0.519–0.569)	0.555 (0.530–0.580)
$H_{t\leq 1y}$	0.540 (0.514–0.567)	0.540 (0.515–0.565)	0.529 (0.504–0.554)	0.568 (0.544–0.593)	0.565 (0.539–0.589)	0.552 (0.525–0.578)	0.556 (0.530–0.582)
$H_{t\leq 3y}$	0.531 (0.504–0.558)	0.547 (0.520–0.572)	0.548 (0.522–0.572)	0.552 (0.526–0.578)	0.577 (0.552–0.602)	0.548 (0.522–0.573)	0.564 (0.539–0.590)
$H_{t\leq C}$	0.525 (0.500–0.550)	0.540 (0.516–0.564)	0.558 (0.533–0.584)	0.557 (0.531–0.583)	0.574 (0.550–0.599)	0.559 (0.534–0.585)	0.572 (0.547–0.596)
$H_{t\leq C,w=1d}$	0.529 (0.503–0.555)	0.544 (0.518–0.571)	0.556 (0.531–0.583)	0.558 (0.533–0.583)	0.582 (0.558–0.608)	0.571 (0.547–0.596)	0.558 (0.532–0.583)
$H_{t\leq C,w=2d}$	0.525 (0.499–0.550)	0.533 (0.507–0.559)	0.549 (0.524–0.576)	0.559 (0.532–0.584)	0.581 (0.556–0.605)	0.575 (0.552–0.600)	0.566 (0.541–0.591)
T_3: One-year clinical instability after latest hospitalization							
$H_{t=0}$	0.815 (0.800–0.829)	0.829 (0.816–0.842)	0.823 (0.808–0.837)	0.835 (0.821–0.848)	0.846 (0.833–0.858)	0.842 (0.828–0.855)	0.840 (0.827–0.854)
$H_{t\leq 1y}$	0.821 (0.808–0.835)	0.830 (0.815–0.843)	0.823 (0.809–0.837)	0.841 (0.828–0.854)	0.859 (0.846–0.870)	0.849 (0.834–0.861)	0.852 (0.838–0.864)
$H_{t\leq 3y}$	0.810 (0.796–0.824)	0.838 (0.825–0.851)	0.835 (0.821–0.849)	0.841 (0.829–0.854)	0.854 (0.841–0.866)	0.848 (0.835–0.861)	0.850 (0.837–0.862)
$H_{t\leq C}$	0.815 (0.802–0.829)	0.835 (0.821–0.849)	0.836 (0.822–0.849)	0.843 (0.830–0.856)	0.845 (0.831–0.858)	0.851 (0.838–0.864)	0.840 (0.826–0.855)
$H_{t\leq C,w=1d}$	0.805 (0.789–0.819)	0.837 (0.824–0.851)	0.851 (0.839–0.864)	0.843 (0.830–0.856)	0.854 (0.842–0.865)	0.855 (0.842–0.867)	0.854 (0.842–0.865)
$H_{t\leq C,w=2d}$	0.807 (0.792–0.822)	0.828 (0.815–0.841)	0.839 (0.826–0.853)	0.841 (0.827–0.853)	0.860 (0.849–0.872)	0.851 (0.838–0.863)	0.854 (0.841–0.866)

Table S11: Ablation of the history H across three clinical tasks T evaluated by AUROC (↑) for Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.

History	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx							
$H_{t=0}$	0.638 (0.625–0.652)	0.654 (0.640–0.668)	0.652 (0.638–0.665)	0.655 (0.641–0.667)	0.666 (0.653–0.679)	0.649 (0.636–0.663)	0.656 (0.642–0.669)
$H_{t\leq 1y}$	0.638 (0.624–0.652)	0.652 (0.639–0.666)	0.655 (0.641–0.668)	0.661 (0.647–0.675)	0.673 (0.659–0.686)	0.653 (0.639–0.666)	0.662 (0.650–0.676)
$H_{t\leq 3y}$	0.644 (0.629–0.658)	0.650 (0.637–0.664)	0.658 (0.644–0.672)	0.661 (0.647–0.674)	0.669 (0.655–0.682)	0.653 (0.640–0.667)	0.661 (0.648–0.675)
$H_{t\leq C}$	0.646 (0.632–0.659)	0.650 (0.636–0.663)	0.651 (0.637–0.664)	0.656 (0.643–0.670)	0.673 (0.660–0.686)	0.657 (0.644–0.671)	0.666 (0.653–0.680)
$H_{t\leq C,w=1d}$	0.644 (0.630–0.658)	0.652 (0.638–0.666)	0.661 (0.647–0.675)	0.657 (0.643–0.671)	0.665 (0.651–0.678)	0.668 (0.654–0.681)	0.661 (0.647–0.674)
$H_{t\leq C,w=2d}$	0.645 (0.630–0.658)	0.643 (0.629–0.657)	0.662 (0.649–0.675)	0.665 (0.651–0.679)	0.669 (0.655–0.681)	0.667 (0.654–0.681)	0.660 (0.647–0.673)
T_2: One-year mortality after initial HF Dx							
$H_{t=0}$	0.780 (0.767–0.792)	0.785 (0.774–0.797)	0.796 (0.784–0.808)	0.794 (0.782–0.805)	0.803 (0.793–0.816)	0.790 (0.779–0.802)	0.791 (0.779–0.803)
$H_{t\leq 1y}$	0.785 (0.773–0.798)	0.787 (0.776–0.799)	0.783 (0.771–0.794)	0.798 (0.786–0.810)	0.801 (0.789–0.812)	0.793 (0.781–0.805)	0.792 (0.780–0.804)
$H_{t\leq 3y}$	0.783 (0.770–0.795)	0.790 (0.779–0.802)	0.791 (0.779–0.803)	0.792 (0.780–0.803)	0.803 (0.792–0.816)	0.794 (0.783–0.806)	0.797 (0.785–0.809)
$H_{t\leq C}$	0.778 (0.765–0.790)	0.787 (0.775–0.799)	0.795 (0.783–0.808)	0.793 (0.781–0.805)	0.805 (0.794–0.817)	0.799 (0.787–0.811)	0.801 (0.790–0.813)
$H_{t\leq C,w=1d}$	0.779 (0.767–0.792)	0.790 (0.778–0.802)	0.798 (0.787–0.811)	0.793 (0.781–0.805)	0.806 (0.795–0.818)	0.802 (0.791–0.814)	0.799 (0.788–0.811)
$H_{t\leq C,w=2d}$	0.778 (0.766–0.791)	0.779 (0.767–0.791)	0.796 (0.783–0.808)	0.793 (0.781–0.805)	0.808 (0.797–0.819)	0.800 (0.788–0.812)	0.798 (0.786–0.809)
T_3: One-year clinical instability after latest hospitalization							
$H_{t=0}$	0.848 (0.838–0.857)	0.857 (0.848–0.865)	0.851 (0.843–0.860)	0.863 (0.855–0.871)	0.873 (0.865–0.881)	0.869 (0.861–0.878)	0.867 (0.859–0.875)
$H_{t\leq 1y}$	0.854 (0.845–0.863)	0.860 (0.852–0.869)	0.855 (0.846–0.864)	0.867 (0.859–0.876)	0.882 (0.874–0.890)	0.875 (0.867–0.883)	0.878 (0.870–0.886)
$H_{t\leq 3y}$	0.847 (0.837–0.855)	0.865 (0.857–0.873)	0.865 (0.856–0.873)	0.869 (0.860–0.877)	0.880 (0.872–0.888)	0.875 (0.867–0.883)	0.877 (0.868–0.885)
$H_{t\leq C}$	0.848 (0.839–0.857)	0.866 (0.858–0.874)	0.863 (0.854–0.871)	0.871 (0.863–0.879)	0.874 (0.866–0.882)	0.875 (0.868–0.883)	0.872 (0.864–0.881)
$H_{t\leq C,w=1d}$	0.843 (0.834–0.852)	0.869 (0.860–0.877)	0.877 (0.869–0.885)	0.871 (0.863–0.879)	0.879 (0.871–0.887)	0.881 (0.873–0.888)	0.879 (0.871–0.886)
$H_{t\leq C,w=2d}$	0.844 (0.835–0.854)	0.861 (0.852–0.870)	0.868 (0.860–0.877)	0.870 (0.862–0.879)	0.886 (0.879–0.894)	0.877 (0.869–0.885)	0.879 (0.871–0.888)

Table S12: Ablation of the history H across three clinical tasks T evaluated by Brier score (↓) for Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.

History	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx							
$H_{t=0}$	0.229 (0.225–0.233)	0.224 (0.220–0.227)	0.224 (0.220–0.227)	0.223 (0.220–0.227)	0.220 (0.217–0.224)	0.226 (0.222–0.230)	0.223 (0.220–0.226)
$H_{t\leq 1y}$	0.229 (0.225–0.234)	0.224 (0.220–0.228)	0.223 (0.219–0.226)	0.222 (0.218–0.226)	0.219 (0.215–0.222)	0.225 (0.220–0.229)	0.223 (0.219–0.227)
$H_{t\leq 3y}$	0.228 (0.224–0.233)	0.224 (0.220–0.227)	0.223 (0.219–0.227)	0.222 (0.218–0.226)	0.222 (0.218–0.227)	0.225 (0.221–0.229)	0.222 (0.219–0.225)
$H_{t\leq C}$	0.229 (0.224–0.233)	0.225 (0.221–0.229)	0.224 (0.221–0.228)	0.223 (0.219–0.226)	0.219 (0.215–0.223)	0.223 (0.219–0.227)	0.221 (0.218–0.224)
$H_{t\leq C,w=1d}$	0.229 (0.225–0.234)	0.226 (0.222–0.230)	0.222 (0.218–0.226)	0.222 (0.218–0.226)	0.222 (0.218–0.227)	0.223 (0.218–0.228)	0.225 (0.220–0.229)
$H_{t\leq C,w=2d}$	0.229 (0.225–0.234)	0.228 (0.224–0.232)	0.221 (0.218–0.225)	0.222 (0.218–0.226)	0.219 (0.216–0.223)	0.223 (0.218–0.227)	0.223 (0.219–0.228)
T_2: One-year mortality after initial HF Dx							
$H_{t=0}$	0.152 (0.147–0.158)	0.152 (0.147–0.156)	0.147 (0.142–0.152)	0.149 (0.143–0.154)	0.146 (0.140–0.151)	0.150 (0.145–0.155)	0.150 (0.144–0.155)
$H_{t\leq 1y}$	0.151 (0.146–0.156)	0.152 (0.146–0.157)	0.155 (0.150–0.159)	0.148 (0.142–0.153)	0.146 (0.142–0.151)	0.150 (0.144–0.156)	0.149 (0.143–0.154)
$H_{t\leq 3y}$	0.152 (0.147–0.157)	0.149 (0.144–0.154)	0.150 (0.145–0.155)	0.149 (0.144–0.154)	0.145 (0.140–0.151)	0.150 (0.145–0.155)	0.147 (0.142–0.153)
$H_{t\leq C}$	0.153 (0.148–0.159)	0.151 (0.145–0.156)	0.148 (0.142–0.153)	0.148 (0.143–0.154)	0.146 (0.140–0.151)	0.148 (0.142–0.154)	0.147 (0.141–0.153)
$H_{t\leq C,w=1d}$	0.153 (0.148–0.158)	0.150 (0.144–0.155)	0.148 (0.143–0.153)	0.148 (0.142–0.153)	0.144 (0.138–0.149)	0.145 (0.140–0.150)	0.146 (0.141–0.151)
$H_{t\leq C,w=2d}$	0.153 (0.147–0.159)	0.152 (0.147–0.158)	0.148 (0.143–0.153)	0.148 (0.142–0.153)	0.144 (0.139–0.150)	0.145 (0.140–0.150)	0.146 (0.141–0.151)
T_3: One-year clinical instability after latest hospitalization							
$H_{t=0}$	0.159 (0.154–0.164)	0.157 (0.152–0.162)	0.160 (0.154–0.164)	0.150 (0.146–0.155)	0.149 (0.144–0.155)	0.148 (0.143–0.153)	0.150 (0.144–0.155)
$H_{t\leq 1y}$	0.155 (0.151–0.161)	0.154 (0.150–0.158)	0.159 (0.153–0.164)	0.151 (0.145–0.156)	0.145 (0.140–0.151)	0.144 (0.139–0.149)	0.141 (0.136–0.147)
$H_{t\leq 3y}$	0.160 (0.155–0.165)	0.151 (0.146–0.156)	0.151 (0.146–0.156)	0.149 (0.143–0.155)	0.144 (0.138–0.150)	0.145 (0.141–0.150)	0.143 (0.137–0.148)
$H_{t\leq C}$	0.159 (0.154–0.164)	0.150 (0.145–0.156)	0.151 (0.146–0.156)	0.147 (0.141–0.152)	0.145 (0.140–0.151)	0.145 (0.140–0.150)	0.146 (0.141–0.152)
$H_{t\leq C,w=1d}$	0.162 (0.156–0.167)	0.147 (0.142–0.153)	0.148 (0.143–0.154)	0.148 (0.143–0.153)	0.141 (0.136–0.146)	0.140 (0.135–0.145)	0.142 (0.138–0.146)
$H_{t\leq C,w=2d}$	0.160 (0.155–0.166)	0.152 (0.147–0.157)	0.152 (0.146–0.157)	0.149 (0.144–0.153)	0.147 (0.142–0.152)	0.142 (0.137–0.147)	0.144 (0.139–0.150)

Table S13: Ablation of incremental concept augmentation across three clinical tasks T evaluated by AUPRC (↑) for Medium-sized sequence models and $C = 512$. Here, feature group F_1 refers to DX, F_2 to DX+VIT, F_3 to DX+VIT+LAB, F_4 to DX+VIT+LAB+MED and F_5 to F_4 inclusive PRO. Gray background highlights common setup shared across all ablations.

Feature	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx							
F_1	0.487 (0.467–0.507)	0.502 (0.483–0.523)	0.518 (0.498–0.538)	0.510 (0.490–0.530)	0.518 (0.498–0.538)	0.514 (0.494–0.534)	0.520 (0.500–0.540)
F_2	0.498 (0.478–0.519)	0.513 (0.492–0.534)	0.518 (0.497–0.538)	0.517 (0.497–0.538)	0.537 (0.516–0.557)	0.526 (0.505–0.546)	0.531 (0.509–0.551)
F_3	0.506 (0.486–0.526)	0.521 (0.500–0.541)	0.523 (0.502–0.545)	0.512 (0.491–0.532)	0.527 (0.507–0.549)	0.519 (0.498–0.540)	0.521 (0.502–0.542)
F_4	0.515 (0.494–0.534)	0.537 (0.517–0.557)	0.543 (0.523–0.564)	0.537 (0.516–0.558)	0.553 (0.533–0.574)	0.538 (0.516–0.558)	0.552 (0.533–0.572)
ALL	0.519 (0.497–0.540)	0.535 (0.516–0.556)	0.529 (0.508–0.549)	0.534 (0.513–0.556)	0.555 (0.535–0.575)	0.543 (0.523–0.565)	0.542 (0.521–0.562)
T_2: One-year mortality after initial HF Dx							
F_1	0.495 (0.471–0.521)	0.482 (0.456–0.506)	0.499 (0.473–0.525)	0.491 (0.466–0.517)	0.507 (0.481–0.533)	0.500 (0.473–0.525)	0.497 (0.471–0.524)
F_2	0.522 (0.497–0.545)	0.519 (0.494–0.544)	0.522 (0.497–0.548)	0.530 (0.504–0.555)	0.538 (0.512–0.564)	0.530 (0.503–0.557)	0.521 (0.497–0.546)
F_3	0.518 (0.492–0.543)	0.529 (0.503–0.553)	0.538 (0.512–0.562)	0.551 (0.527–0.575)	0.549 (0.523–0.574)	0.523 (0.498–0.550)	0.553 (0.528–0.579)
F_4	0.511 (0.486–0.535)	0.542 (0.517–0.569)	0.555 (0.529–0.580)	0.549 (0.523–0.575)	0.569 (0.544–0.595)	0.558 (0.533–0.583)	0.561 (0.536–0.587)
ALL	0.525 (0.500–0.550)	0.540 (0.516–0.564)	0.558 (0.533–0.584)	0.557 (0.531–0.583)	0.574 (0.550–0.599)	0.559 (0.534–0.585)	0.572 (0.547–0.596)
T_3: One-year clinical instability after latest hospitalization							
F_1	0.773 (0.756–0.789)	0.785 (0.770–0.802)	0.784 (0.769–0.800)	0.794 (0.779–0.809)	0.815 (0.800–0.829)	0.809 (0.794–0.823)	0.805 (0.790–0.819)
F_2	0.784 (0.769–0.800)	0.813 (0.798–0.828)	0.819 (0.805–0.833)	0.818 (0.804–0.832)	0.832 (0.818–0.846)	0.831 (0.818–0.844)	0.816 (0.801–0.831)
F_3	0.794 (0.779–0.809)	0.824 (0.810–0.838)	0.844 (0.831–0.857)	0.833 (0.819–0.846)	0.847 (0.834–0.859)	0.844 (0.830–0.856)	0.829 (0.815–0.843)
F_4	0.807 (0.792–0.822)	0.839 (0.826–0.852)	0.849 (0.837–0.861)	0.845 (0.833–0.857)	0.855 (0.842–0.867)	0.845 (0.832–0.857)	0.849 (0.836–0.861)
ALL	0.815 (0.802–0.829)	0.835 (0.821–0.849)	0.836 (0.822–0.849)	0.843 (0.830–0.856)	0.845 (0.831–0.858)	0.851 (0.838–0.864)	0.840 (0.826–0.855)

Table S14: Ablation of incremental concept augmentation across three clinical tasks T evaluated by AUROC (↑) for Medium-sized sequence models and $C = 512$. Here, F_1 refers to DX, F_2 to DX+VIT, F_3 to DX+VIT+LAB, F_4 to DX+VIT+LAB+MED and F_5 to F_4 inclusive PRO. Gray background highlights common setup shared across all ablations.

Feature	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx							
F_1	0.619 (0.605–0.633)	0.624 (0.610–0.638)	0.637 (0.623–0.649)	0.631 (0.617–0.645)	0.640 (0.626–0.654)	0.639 (0.626–0.652)	0.640 (0.625–0.653)
F_2	0.628 (0.614–0.642)	0.639 (0.624–0.653)	0.645 (0.631–0.658)	0.641 (0.626–0.654)	0.655 (0.642–0.668)	0.649 (0.635–0.662)	0.651 (0.636–0.664)
F_3	0.633 (0.619–0.646)	0.644 (0.631–0.658)	0.644 (0.630–0.658)	0.633 (0.618–0.646)	0.650 (0.637–0.664)	0.639 (0.625–0.654)	0.646 (0.631–0.659)
F_4	0.636 (0.623–0.650)	0.658 (0.644–0.672)	0.659 (0.645–0.673)	0.657 (0.642–0.671)	0.672 (0.659–0.686)	0.660 (0.647–0.674)	0.669 (0.655–0.682)
ALL	0.646 (0.632–0.659)	0.650 (0.636–0.663)	0.651 (0.637–0.664)	0.656 (0.643–0.670)	0.673 (0.660–0.686)	0.657 (0.644–0.671)	0.666 (0.653–0.680)
T_2: One-year mortality after initial HF Dx							
F_1	0.756 (0.744–0.769)	0.748 (0.735–0.761)	0.756 (0.743–0.770)	0.752 (0.738–0.765)	0.766 (0.753–0.779)	0.761 (0.747–0.773)	0.755 (0.742–0.767)
F_2	0.763 (0.751–0.776)	0.772 (0.760–0.784)	0.774 (0.762–0.786)	0.772 (0.760–0.786)	0.783 (0.771–0.795)	0.778 (0.766–0.791)	0.774 (0.762–0.786)
F_3	0.772 (0.759–0.785)	0.779 (0.767–0.792)	0.782 (0.770–0.795)	0.784 (0.771–0.796)	0.789 (0.777–0.802)	0.782 (0.769–0.794)	0.792 (0.780–0.804)
F_4	0.776 (0.764–0.788)	0.786 (0.774–0.798)	0.795 (0.784–0.807)	0.789 (0.777–0.801)	0.802 (0.791–0.813)	0.799 (0.788–0.811)	0.798 (0.786–0.810)
ALL	0.778 (0.765–0.790)	0.787 (0.775–0.799)	0.795 (0.783–0.808)	0.793 (0.781–0.805)	0.805 (0.794–0.817)	0.799 (0.787–0.811)	0.801 (0.790–0.813)
T_3: One-year clinical instability after latest hospitalization							
F_1	0.813 (0.802–0.824)	0.823 (0.813–0.833)	0.822 (0.812–0.832)	0.828 (0.818–0.838)	0.848 (0.839–0.857)	0.845 (0.835–0.853)	0.838 (0.828–0.846)
F_2	0.826 (0.816–0.836)	0.850 (0.841–0.859)	0.858 (0.849–0.866)	0.854 (0.845–0.862)	0.864 (0.856–0.872)	0.862 (0.853–0.870)	0.855 (0.846–0.864)
F_3	0.834 (0.825–0.844)	0.857 (0.848–0.865)	0.873 (0.865–0.880)	0.863 (0.854–0.871)	0.874 (0.865–0.882)	0.872 (0.863–0.879)	0.862 (0.853–0.870)
F_4	0.845 (0.835–0.854)	0.869 (0.860–0.877)	0.874 (0.866–0.882)	0.870 (0.862–0.878)	0.879 (0.871–0.887)	0.873 (0.866–0.881)	0.875 (0.867–0.883)
ALL	0.848 (0.839–0.857)	0.866 (0.858–0.874)	0.863 (0.854–0.871)	0.871 (0.863–0.879)	0.874 (0.866–0.882)	0.875 (0.868–0.883)	0.872 (0.864–0.881)

Table S15: Ablation of incremental concept augmentation across three clinical tasks T evaluated by Brier score (↓) for Medium-sized sequence models and $C = 512$. Here, F_1 refers to DX, F_2 to DX+VIT, F_3 to DX+VIT+LAB, F_4 to DX+VIT+LAB+MED and F_5 to F_4 inclusive PRO. Gray background highlights common setup shared across all ablations.

Feature	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx							
F_1	0.232 (0.228–0.236)	0.228 (0.225–0.232)	0.227 (0.223–0.230)	0.228 (0.225–0.230)	0.226 (0.223–0.230)	0.226 (0.222–0.230)	0.227 (0.224–0.230)
F_2	0.231 (0.227–0.235)	0.225 (0.222–0.229)	0.225 (0.221–0.228)	0.226 (0.223–0.230)	0.225 (0.220–0.229)	0.224 (0.220–0.228)	0.223 (0.220–0.227)
F_3	0.230 (0.226–0.234)	0.224 (0.221–0.228)	0.225 (0.222–0.228)	0.228 (0.224–0.232)	0.223 (0.220–0.227)	0.227 (0.224–0.231)	0.226 (0.222–0.229)
F_4	0.230 (0.226–0.235)	0.222 (0.218–0.226)	0.222 (0.218–0.227)	0.223 (0.219–0.227)	0.219 (0.215–0.222)	0.221 (0.218–0.225)	0.220 (0.216–0.224)
ALL	0.229 (0.224–0.233)	0.225 (0.221–0.229)	0.224 (0.221–0.228)	0.223 (0.219–0.226)	0.219 (0.215–0.223)	0.223 (0.219–0.227)	0.221 (0.218–0.224)
T_2: One-year mortality after initial HF Dx							
F_1	0.158 (0.153–0.164)	0.161 (0.156–0.166)	0.158 (0.153–0.163)	0.159 (0.154–0.164)	0.156 (0.151–0.162)	0.157 (0.152–0.162)	0.162 (0.158–0.167)
F_2	0.156 (0.150–0.161)	0.155 (0.150–0.160)	0.154 (0.148–0.159)	0.153 (0.148–0.159)	0.151 (0.146–0.157)	0.152 (0.147–0.157)	0.157 (0.152–0.161)
F_3	0.155 (0.150–0.161)	0.152 (0.147–0.158)	0.151 (0.146–0.156)	0.150 (0.145–0.155)	0.151 (0.146–0.157)	0.154 (0.148–0.159)	0.148 (0.143–0.153)
F_4	0.155 (0.150–0.160)	0.153 (0.148–0.159)	0.151 (0.145–0.156)	0.152 (0.148–0.156)	0.146 (0.141–0.152)	0.148 (0.142–0.153)	0.147 (0.142–0.151)
ALL	0.153 (0.148–0.159)	0.151 (0.145–0.156)	0.148 (0.142–0.153)	0.148 (0.143–0.154)	0.146 (0.140–0.151)	0.148 (0.142–0.154)	0.147 (0.141–0.153)
T_3: One-year clinical instability after latest hospitalization							
F_1	0.176 (0.171–0.181)	0.172 (0.167–0.177)	0.172 (0.167–0.177)	0.174 (0.169–0.179)	0.159 (0.154–0.164)	0.162 (0.157–0.167)	0.165 (0.160–0.171)
F_2	0.170 (0.165–0.175)	0.159 (0.154–0.165)	0.153 (0.148–0.158)	0.156 (0.151–0.161)	0.150 (0.145–0.155)	0.152 (0.147–0.157)	0.155 (0.150–0.160)
F_3	0.166 (0.161–0.171)	0.154 (0.149–0.159)	0.145 (0.140–0.151)	0.152 (0.147–0.157)	0.147 (0.142–0.152)	0.146 (0.141–0.151)	0.151 (0.146–0.156)
F_4	0.160 (0.155–0.166)	0.150 (0.144–0.155)	0.145 (0.140–0.150)	0.147 (0.143–0.152)	0.142 (0.137–0.146)	0.144 (0.139–0.149)	0.143 (0.138–0.148)
ALL	0.159 (0.154–0.164)	0.150 (0.145–0.156)	0.151 (0.146–0.156)	0.147 (0.141–0.152)	0.145 (0.140–0.151)	0.145 (0.140–0.150)	0.146 (0.141–0.152)

Table S16: Ablation of varying training set ratios across three clinical tasks T evaluated by AUPRC (↑) for Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.

Ratio	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx							
25%	0.502 (0.482–0.522)	0.516 (0.496–0.537)	0.523 (0.503–0.544)	0.522 (0.501–0.541)	0.521 (0.501–0.542)	0.528 (0.508–0.550)	0.514 (0.493–0.534)
50%	0.510 (0.492–0.531)	0.531 (0.511–0.551)	0.522 (0.502–0.542)	0.535 (0.515–0.556)	0.545 (0.525–0.566)	0.526 (0.506–0.546)	0.543 (0.521–0.562)
75%	0.515 (0.495–0.536)	0.527 (0.507–0.546)	0.529 (0.508–0.549)	0.539 (0.518–0.560)	0.547 (0.526–0.568)	0.523 (0.503–0.545)	0.541 (0.521–0.561)
100%	0.519 (0.497–0.540)	0.535 (0.516–0.556)	0.529 (0.508–0.549)	0.534 (0.513–0.556)	0.555 (0.535–0.575)	0.543 (0.523–0.565)	0.542 (0.521–0.562)
T_2: One-year mortality after initial HF Dx							
25%	0.506 (0.481–0.533)	0.473 (0.448–0.499)	0.475 (0.450–0.500)	0.521 (0.497–0.547)	0.518 (0.493–0.544)	0.510 (0.485–0.535)	0.483 (0.457–0.507)
50%	0.508 (0.481–0.534)	0.523 (0.498–0.548)	0.521 (0.496–0.546)	0.534 (0.508–0.561)	0.564 (0.539–0.589)	0.539 (0.511–0.563)	0.528 (0.503–0.553)
75%	0.516 (0.492–0.541)	0.525 (0.499–0.550)	0.539 (0.513–0.565)	0.545 (0.520–0.571)	0.574 (0.548–0.600)	0.552 (0.525–0.577)	0.553 (0.529–0.577)
100%	0.525 (0.500–0.550)	0.540 (0.516–0.564)	0.558 (0.533–0.584)	0.557 (0.531–0.583)	0.574 (0.550–0.599)	0.559 (0.534–0.585)	0.572 (0.547–0.596)
T_3: One-year clinical instability after latest hospitalization							
25%	0.794 (0.778–0.810)	0.804 (0.789–0.819)	0.791 (0.775–0.807)	0.805 (0.789–0.819)	0.829 (0.815–0.842)	0.808 (0.792–0.822)	0.810 (0.794–0.824)
50%	0.807 (0.793–0.820)	0.821 (0.805–0.836)	0.832 (0.818–0.846)	0.829 (0.814–0.843)	0.840 (0.826–0.853)	0.838 (0.824–0.851)	0.841 (0.828–0.854)
75%	0.809 (0.795–0.823)	0.831 (0.817–0.844)	0.841 (0.828–0.854)	0.836 (0.822–0.849)	0.845 (0.833–0.858)	0.844 (0.831–0.857)	0.837 (0.824–0.851)
100%	0.815 (0.802–0.829)	0.835 (0.821–0.849)	0.836 (0.822–0.849)	0.843 (0.830–0.856)	0.845 (0.831–0.858)	0.851 (0.838–0.864)	0.840 (0.826–0.855)

Table S17: Ablation of varying training set ratios across three clinical tasks T evaluated by AUROC (↑) for Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.

Ratio	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx							
25%	0.624 (0.610–0.638)	0.641 (0.628–0.655)	0.637 (0.624–0.651)	0.644 (0.629–0.657)	0.645 (0.631–0.660)	0.646 (0.633–0.660)	0.640 (0.626–0.654)
50%	0.631 (0.617–0.644)	0.645 (0.631–0.659)	0.641 (0.627–0.656)	0.652 (0.639–0.666)	0.663 (0.650–0.677)	0.646 (0.632–0.660)	0.663 (0.649–0.676)
75%	0.639 (0.625–0.653)	0.647 (0.634–0.661)	0.648 (0.634–0.662)	0.659 (0.645–0.672)	0.666 (0.653–0.679)	0.647 (0.635–0.661)	0.663 (0.650–0.677)
100%	0.646 (0.632–0.659)	0.650 (0.636–0.663)	0.651 (0.637–0.664)	0.656 (0.643–0.670)	0.673 (0.660–0.686)	0.657 (0.644–0.671)	0.666 (0.653–0.680)
T_2: One-year mortality after initial HF Dx							
25%	0.759 (0.746–0.772)	0.743 (0.729–0.756)	0.751 (0.738–0.764)	0.769 (0.757–0.782)	0.777 (0.765–0.790)	0.767 (0.755–0.780)	0.759 (0.746–0.771)
50%	0.771 (0.757–0.783)	0.774 (0.762–0.786)	0.779 (0.767–0.791)	0.781 (0.770–0.794)	0.796 (0.784–0.808)	0.786 (0.774–0.798)	0.781 (0.769–0.793)
75%	0.770 (0.757–0.783)	0.776 (0.763–0.789)	0.783 (0.771–0.796)	0.788 (0.777–0.801)	0.803 (0.791–0.815)	0.790 (0.778–0.802)	0.790 (0.778–0.802)
100%	0.778 (0.765–0.790)	0.787 (0.775–0.799)	0.795 (0.783–0.808)	0.793 (0.781–0.805)	0.805 (0.794–0.817)	0.799 (0.787–0.811)	0.801 (0.790–0.813)
T_3: One-year clinical instability after latest hospitalization							
25%	0.834 (0.825–0.844)	0.840 (0.831–0.849)	0.832 (0.823–0.842)	0.846 (0.836–0.855)	0.861 (0.853–0.869)	0.842 (0.833–0.851)	0.850 (0.841–0.859)
50%	0.842 (0.833–0.851)	0.857 (0.848–0.865)	0.864 (0.856–0.872)	0.860 (0.852–0.869)	0.873 (0.865–0.881)	0.866 (0.857–0.874)	0.869 (0.861–0.878)
75%	0.842 (0.833–0.851)	0.863 (0.854–0.871)	0.870 (0.862–0.878)	0.866 (0.858–0.874)	0.872 (0.864–0.881)	0.874 (0.865–0.882)	0.867 (0.859–0.875)
100%	0.848 (0.839–0.857)	0.866 (0.858–0.874)	0.863 (0.854–0.871)	0.871 (0.863–0.879)	0.874 (0.866–0.882)	0.875 (0.868–0.883)	0.872 (0.864–0.881)

Table S18: Ablation of varying training set ratios across three clinical tasks T evaluated by Brier score (↓) for Medium-sized sequence models and $C = 512$. Gray background highlights common setup shared across all ablations.

Ratio	XGBoost	BERT	XLNet	ModernBERT	Llama	Mamba	Mamba2
T_1: One-year clinical instability after initial HF Dx							
25%	0.238 (0.233–0.243)	0.225 (0.222–0.229)	0.227 (0.223–0.230)	0.227 (0.223–0.231)	0.225 (0.221–0.230)	0.226 (0.222–0.229)	0.230 (0.225–0.235)
50%	0.234 (0.229–0.238)	0.228 (0.224–0.232)	0.226 (0.223–0.229)	0.225 (0.221–0.229)	0.224 (0.220–0.228)	0.224 (0.221–0.228)	0.221 (0.217–0.224)
75%	0.231 (0.226–0.235)	0.225 (0.221–0.229)	0.223 (0.220–0.227)	0.222 (0.218–0.226)	0.222 (0.218–0.226)	0.227 (0.222–0.231)	0.223 (0.218–0.227)
100%	0.229 (0.224–0.233)	0.225 (0.221–0.229)	0.224 (0.221–0.228)	0.223 (0.219–0.226)	0.219 (0.215–0.223)	0.223 (0.219–0.227)	0.221 (0.218–0.224)
T_2: One-year mortality after initial HF Dx							
25%	0.160 (0.154–0.165)	0.163 (0.158–0.169)	0.159 (0.155–0.164)	0.155 (0.150–0.160)	0.153 (0.149–0.158)	0.157 (0.151–0.162)	0.159 (0.154–0.163)
50%	0.156 (0.151–0.162)	0.153 (0.149–0.158)	0.154 (0.150–0.159)	0.153 (0.147–0.159)	0.149 (0.143–0.155)	0.151 (0.145–0.156)	0.154 (0.148–0.160)
75%	0.156 (0.150–0.161)	0.153 (0.148–0.157)	0.151 (0.146–0.156)	0.151 (0.146–0.156)	0.148 (0.143–0.154)	0.148 (0.143–0.154)	0.153 (0.147–0.159)
100%	0.153 (0.148–0.159)	0.151 (0.145–0.156)	0.148 (0.142–0.153)	0.148 (0.143–0.154)	0.146 (0.140–0.151)	0.148 (0.142–0.154)	0.147 (0.141–0.153)
T_3: One-year clinical instability after latest hospitalization							
25%	0.168 (0.162–0.174)	0.171 (0.165–0.177)	0.167 (0.163–0.171)	0.160 (0.156–0.165)	0.152 (0.147–0.157)	0.164 (0.158–0.169)	0.163 (0.157–0.169)
50%	0.162 (0.157–0.168)	0.156 (0.151–0.160)	0.153 (0.148–0.157)	0.153 (0.147–0.158)	0.144 (0.139–0.149)	0.159 (0.153–0.165)	0.150 (0.145–0.154)
75%	0.162 (0.157–0.168)	0.151 (0.145–0.156)	0.146 (0.142–0.151)	0.152 (0.147–0.157)	0.146 (0.140–0.151)	0.146 (0.141–0.150)	0.150 (0.145–0.154)
100%	0.159 (0.154–0.164)	0.150 (0.145–0.156)	0.151 (0.146–0.156)	0.147 (0.141–0.152)	0.145 (0.140–0.151)	0.145 (0.140–0.150)	0.146 (0.141–0.152)