CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks

Chao Pang CP3016@CUMC.COLUMBIA.EDU
Xinzhuo Jiang XJ2193@CUMC.COLUMBIA.EDU
Krishna S. Kalluri KK3326@CUMC.COLUMBIA.EDU
Matthew Spotnitz MES2165@CUMC.COLUMBIA.EDU
Columbia University Irving Medical Center

RuiJun Chen RUIJUN.CHEN@GMAIL.COM
Geisinger

Adler Perotte AJP2120@CUMC.COLUMBIA.EDU
Karthik Natarajan KN2174@CUMC.COLUMBIA.EDU
Columbia University Irving Medical Center

Abstract

Embedding algorithms are increasingly used to represent clinical concepts in healthcare for improving machine learning tasks such as clinical phenotyping and disease prediction. Recent studies have adapted state-of-the-art bidirectional encoder representations from transformers (BERT) architecture to structured electronic health records (EHR) data for the generation of contextualized concept embeddings, yet do not fully incorporate temporal data across multiple clinical domains. Therefore we developed a new BERT adaptation, CEHR-BERT, to incorporate temporal information using a hybrid approach by augmenting the input to BERT using artificial time tokens, incorporating time, age, and concept embeddings, and introducing a new second learning objective for visit type. CEHR-BERT was trained on a subset of clinical data from Columbia University Irving Medical Center-New York Presbyterian Hospital, which includes 2.4M patients, spanning over three decades, and tested using 4-fold evaluation on the following prediction tasks: hospitalization, death, new heart failure (HF) diagnosis, and HF readmission. Our experiments show that CEHR-BERT outperformed existing state-of-the-art clinical BERT adaptations and baseline models across all 4 prediction tasks in both ROC-AUC and PR-AUC. CEHR-BERT also demonstrated strong few-shot learning capability, as our model trained on only 5% of data outperformed comparison models trained on the entire data set. Ablation studies to better understand the contribution of each time component showed incremental gains with every element, suggesting that CEHR-BERT's incorporation of artificial time tokens, time/age embeddings with concept embeddings, and the addition of the second learning objective represents a promising approach for future BERT-based clinical embeddings.

Keywords: Representation learning, Electronic Health Records, Pre-training

1. Introduction

Embedding algorithms, widely used for obtaining low dimensional vector representations of words in natural language processing (NLP) applications, have been increasingly adapted for the representation of clinical concepts in healthcare to improve the development of clinical phenotypes and disease prediction (Glicksberg et al., 2018). Recent advances in contextualized representations such as Bidirectional encoder representations from transformers (BERT) have revolutionized the NLP field, achieving state of the art performance on all benchmark tasks (Devlin et al., 2019; Peters et al., 2018). However, there have been few efforts to apply BERT to structured electronic health record (EHR) data for generating contextualized concept embeddings despite promising results from early BERT adaptations demonstrating improved performance compared to classic embedding algorithms such as word2vec and GloVe (Rasmy et al., 2021; Li et al., 2020).

Despite the differences between structured EHR and text data, a common practice used in the afore-

mentioned BERT adaptations and other classic embedding algorithms is to treat a patient's medical history as a text document where medical concepts are treated as words and ordered chronologically (Beam et al., 2020; Choi et al., 2016; Xiang et al., 2019). Although this representation could capture the rich contextual information of a patient's medical history, the temporal intervals between medical concepts or visits are not preserved; as a consequence, BERT models trained using this patient representation cannot fully leverage temporal information, limiting their performance in downstream prediction tasks. Another challenge is that BERT's second learning objective – Next Sentence Prediction (NSP) (Devlin et al., 2019) does not apply in the context of structured EHR data as the entire patient history is treated as a single sentence. A common approach adopted by others is to disable NSP and pre-train BERT only with Masked Language Modeling (MLM) (Li et al., 2020), yet there is abundant information in structured EHR data that could be leveraged for designing a new secondary learning objective to improve BERT's performance in downstream prediction tasks.

In this paper, we focus on adapting the original BERT architecture for structured EHR data in order to improve disease predictions. We propose a new BERT architecture called CEHR-BERT, where we combine two approaches for encoding temporal information of the structured EHR data by: 1) modifying the patient sequence representation through the insertion of artificial tokens between visits to indicate the time intervals; 2) concatenating both age embeddings and time embeddings to concept embeddings to form temporal concept embeddings. Additionally, we designed a second learning objective – Visit Type Prediction (VTP) for CEHR-BERT that leverages heterogeneous EHR data to further boost the performance of BERT.

2. Related work

A small number of recent studies have sought to adapt BERT for structured EHR data and demonstrated significant performance improvements in their respective evaluations. However, these studies were often limited to a single clinical domain, single visit, or limited in their consideration of time, without fully utilizing the richness of a patient's full medical history.

Li et al. (2020) described the first BERT adaptation for structured EHR data named BEHRT, which pioneered the idea of utilizing multiple types of embeddings to represent patient history, including concept embeddings, visit segment embeddings, age embeddings, and positional embeddings. In addition, the authors inserted **SEP** tokens between visits to indicate the boundaries of visits and enable the execution of BERT as-is on structured EHR data without any modification of the encoder. However, this study only included diagnosis codes in a patient sequence and excluded other clinical domains such as procedures and medications that contain valuable contextual information. In addition, their evaluation focused on diagnosis code prediction instead of disease prediction based on phenotypes.

Another BERT adaptation named G-BERT (Shang et al., 2019) extended BERT to incorporate a graph neural network (GNN). The key idea was to leverage prior knowledge from well crafted medical ontologies to guide the learning of concept embeddings. However, G-BERT was only tailored to medication recommendation; its input data was limited to single visits, and the dataset used for pre-training was relatively small, containing 20K patients.

The latest BERT adaptation, MedBert (Rasmy et al., 2021), had a similar patient representation as BEHRT except that it did not include any temporal information in their model. MedBert excluded age embeddings and visit segment embeddings, along with excluding the **SEP** token inserted between visits in favor of including more concept codes. MedBert was trained on EHR data from 20 million patients and introduced a new second learning objective to predict whether the patient had a prolonged length of stay (defined as inpatient visit longer than 7 days). The authors fine-tuned for two disease prediction tasks using a different data source to demonstrate the potential of transfer learning and showed improved performance. Though a large training data set was used and the study showed improvement in the prediction tasks, the lack of temporal information may have limited its potential performance.

Finally, there are two studies that attempted to incorporate temporal information from structured EHR data (Peng et al., 2019; Che et al., 2018). These studies adopted a similar strategy of incorporating time intervals between neighboring clinical events into their models (e.g. two neighboring visits or lab values in time-series data). In this work, we take a different approach of incorporating time with the introduction of artificial time tokens in CHER-BERT, which will be described in the following section.

3. Data and Preprocessing

3.1. Data

EHR data from Columbia University Irving Medical Center-New York Presbyterian Hospital (CUIMC-NYP) was converted into Observational Medical Outcomes Partnership (OMOP), a common data model used to support observational studies and managed by the Observational Health Data Science and Informatics (OHDSI) open-science community (Hripcsak et al., 2015). The CUIMC-NYP OMOP instance includes numerous data and clinical domains, including visits, conditions, procedures, medications, lab tests, vital signs, and problem lists, among others. Data spans from the early 1980s to present day. We used the CUIMC-NYP OMOP to generate training data and downstream prediction cohorts. To pre-train and fine-tune BERTs, we limited the data to three OMOP domains - conditions, procedures, and medications.

3.2. Data processing and patient representation

For each patient, all medical codes were aggregated from three domains and constructed into a sequence chronologically. In order to incorporate temporal information, we inserted an artificial time token (ATT) between two neighboring visits based on their time interval. The following logic was used for creating ATTs based on the following time intervals between visits: 1) if less than 28 days, ATTs take on the form of W_n where n represents the week number ranging from 0-3 (e.g. W_1); 2) if between 28 days and 365 days, ATTs are in the form of M_n where n represents the month number ranging from 1-11 e.g M_{11}; 3) beyond 365 days then a LT (Long Term) token is inserted. In addition, we added two more special tokens — VS and VE to represent the start and the end of a visit to explicitly define the visit segment, where all the concepts associated with the visit are subsumed by VS and VE. Conceptually, a patient can be represented as a list of visits,

$$\begin{aligned} P = \{&VS,\ v_1,\ VE,\ ATT,\\ &VS,\ v_2,\ VE,\ ATT,\\ &VS,\ v_3,\ VE,\ ATT,\\ &\ldots,\\ &VS,\ v_i,\ VE\} \end{aligned}$$

where v_i represents the ith visit, and each visit consists of a list of medical concepts $v_i = \{c_{i1}, c_{i2}, c_{i3}, \ldots, c_{ij}\}$, see an example in Figure 4. We will refer to EHR patient data representation as patient sequences in the rest of the paper.

4. Methods

Figure 1 shows a high level overview of our adapted BERT architecture. We used multiple sets of embeddings to represent a patient history including concept embeddings, visit segment embeddings, time embeddings and age embeddings. Concept embeddings were used to capture the numeric representations of the concept codes based on underlying co-occurrence statistics, whereas visit segment embeddings were used to indicate the boundaries of visits (values alternating between A and B). Unlike the previous work, we decided to encode both absolute time (time embeddings) and relative time with respect to visits (age embeddings), due to the finding that certain conditions follow a more seasonal pattern (e.g. flu) while other conditions are more age related (e.g. type 2 diabetes). However, because time and age are numeric values that cannot be directly encoded using standard procedures, we therefore followed the methodology proposed by time2vec (Kazemi et al., 2019). A fourier transform was applied to decompose a sequence of time points into a series of sine functions, which are controlled by learnable parameters in order to adapt to specific training data. We concatenated concept, time and age embeddings together, then fed it through a fully connected (FC) layer to bring it back to the original dimension, which became the temporal concept embeddings input for the BERT architecture.

For pre-training, we used the core learning objective MLM and followed the standard procedure described in the original BERT paper (Devlin et al., 2019). In addition, we designed a second learning objective named Visit Type Prediction (VTP) to improve BERT's performance in downstream prediction tasks. VTP was developed based on the observation that different medical concepts are associated with different visit types, and therefore incorporating such domain knowledge may allow BERT to capture additional contextual information. Conceptually, VTP can be thought of as a language translation task, where a sequence of medical concepts are translated to a sequence of visit types (e.g. Inpatient visit v.s. Outpatient visit). To realize this idea, we added a single decoder layer to the BERT architecture to perform VTP. The decoder setup can be summarized as follows: 1) 50% of tokens in the visit sequence are ran-

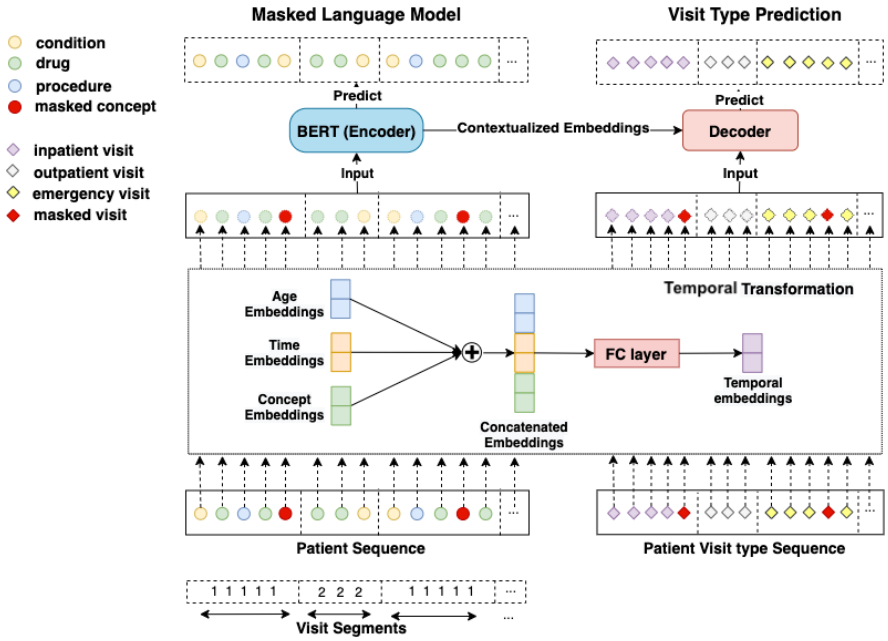

Figure 1: Overview of our BERT architecture on structured EHR data. To distinguish visit boundaries, visit segment embeddings are added to concept embeddings. Next, both visit embeddings and concept embeddings go through a temporal transformation, where concept, age and time embeddings are concatenated together. The concatenated embeddings are then fed into a fully connected layer. This temporal concept embedding becomes the input to BERT. We used the BERT learning objective Masked Language Model as the primary learning objective and introduced an EHR specific secondary learning objective visit type prediction.

domly masked; 2) the visit type sequence undergoes the same temporal transformation as the concept sequence to generate the temporal visit type embeddings; 3) in the decoder, the temporal visit type embeddings and contextualized concept embeddings are combined using multi-headed attention to produce the contextualized visit embeddings; 4) contextualized visit embeddings are used to predict the original visit types for those masked positions in the visit sequence. This second learning objective was trained together with the primary learning objective, MLM. It should be emphasized that the visit sequence is not a list of visits that a patient experienced in the past, but rather, it is a list of visit types constructed from the corresponding concepts in the patient sequence.

5. Experiments and Results

5.1. Experiment setup

CEHR-BERT was pre-trained using EHR data from a Columbia University Irving Medical Center-New York Presbyterian Hospital (CUIMC-NYP). Patients who had at least one visit and more than 5 data points in their medical history were included, resulting in 2.4M patients and 184.7M clinical data points across OMOP domains - condition, procedure, and medication. The data characteristics can be found in Table 1. For pre-training BERT, we used 5 encoders and 8 heads with a dropout rate of 0.1, along with embedding and hidden dimensions of 128. The context window of 300 tokens rather than the standard 512 was used to construct the patient sequence because 300 is enough to capture more than 90% of patients' entire medical histories. For those patients who have a sequence of more than 300 codes, we randomly sliced a subset of their sequence (patient history) for pre-training, while patients with less than 300 codes were post-padded with the **PAD** token. We trained the BERT model for 5 epochs using the Adam optimizer (Kingma and Ba, 2014) with a batch size of 32. The learning rate was initially set to $2e^{-4}$ and the Cosine Annealing LR was used to decay learning

rate after every epoch. All training and testing for this study was done on a Linux based server with 768 GB memory, dual Intel(R) Xeon(R) Gold 6138 CPU @ 2.00GHz processors, and two RTX 2080ti GPUs. The code is available at https://github.com/cumc-dbmi/cehr-bert.

Table 1: Summary statistics of the CUIMC-NYP OMOP instance

	No. of visits per patient	No. of records per patient
mean	14	76
std	29	196
min	1	5
25%	2	10
50%	5	24
75%	14	68
max	1106	31226

5.2. Experiments

To evaluate the model, we followed previously published standards where we fine-tuned BERT together with a Bi-LSTM layer for a set of disease prediction tasks (Rasmy et al., 2021). Then, we conducted a few-shot learning experiment using the same setup except that a subset of the training data was used for fine-tuning and the model performance was evaluated using the full test set (McDermott et al., 2021). In these experiments, we included both BEHRT and MedBert as BERT comparators to better understand the relative performance of our model. For a fair comparison, the BERT comparators were extended to pre-train and fine-tune on three domains including condition, medication, and procedure rather than condition only in their original studies. To understand the contribution of pre-training, we also included a non pre-trained version of CHER-BERT, which will be referred to as R-BERT. In addition, several baseline models were evaluated, including logistic regression (LR), XG-boost, and Bi-LSTM with pre-trained time attention concept embeddings (Xiang et al., 2019). Finally, we conducted ablation studies to understand how different temporal components in our adaption could impact the model performance.

5.2.1. Disease prediction

Disease prediction is the likelihood of a patient experiencing a condition (disease) in a given time window. Table 2 shows the 4 prediction tasks including demographics and patient outcomes. Full definitions can be found in section A. For feature extraction, a one-year observation window was used prior to the entry of the target cohort by default unless stated otherwise in the prediction definition.

Extracting features for sequence models (e.g. LSTM and BERTs) was straightforward, where medical concepts were organized in a sequence using chronological order; and artificial tokens and the **SEP** token were inserted between neighboring visits for CEHR-BERT and BEHRT respectively. The pre-truncate and post-padding strategy was utilized to standardize the size of the inputs for patients who had more/less concepts than the size of the context window. For LR and XG-boost, frequency-based features were constructed by counting the number of occurrences of medical concepts in the observation window. Additionally, medical concepts were rolled up using ontological hierarchies to reduce dimensionality (Ng et al., 2016), for which a detailed explanation could be found in section A.1

For performance evaluation, the 4-fold evaluation was utilized, where in each fold the data was split into three different sets (75:10:15 split for train/val/test). The sequence models were trained for 10 epochs with early stopping to monitor validation loss with the patience set to 1. On the other hand, since frequency based models (LR and XGB) were not fine-tuned for hyper-parameters, 85% of the data was used for training and the remaining 15% was used for testing directly. The python library sklearn was utilized for training with the default configuration. At the end of each fold, the area under the receiver operating characteristics curve (AUC) was calculated using the test set. In addition, PR-AUC (Precision-Recall) was reported because some of the prediction cohorts had imbalanced outcomes e.g. *Discharge Home Death*.

Table 3 shows the average AUC and PR-AUC for all models across 4 prediction tasks, where the best value is highlighted in bold. Overall, sequence models outperformed frequency based models, and BERTs performed better than Bi-LSTM. Among all models evaluated, CEHR-BERT achieved the best performance in both AUC and PR-AUC across all tasks. In particular, CEHR-BERT is the only model that achieved an AUC of 80 in *t2dm hf*, and its PR-AUC exceeded the second best performing model MedBert by more than 10% (from 0.274 to 0.323). The second and third top performing models were MedBert and BEHRT, whose performances were consistently worse

Table 2: Definitions and cohort characteristics of prediction tasks

	HF readmission	*Discharge home Death*	*T2DM HF*	*Hospitalization*
Cohort size	97758	207919	114564	590578
Median age	72	49	61	45
Male	50.30%	33.23%	49.66%	37.46%
Female	49.71%	66.77%	50.34%	62.54%
Outcome	24.16%	4.85%	9.38%	10.90%
Description	30 days all-cause readmission in HF patients (Golas et al., 2018), see A.3 for details.	Mortality within 1 year since discharged to home, see A.4 for details	Life time heart failure prediction since the initial diagnosis of type 2 diabetes mellitus (Rasmy et al., 2021), the medical codes used can be found in A.2	2 year risk of hospitalization starting from the $3rd$ year since the inital entry into the EHR system (Zhang et al., 2018), see A.5 for details

than that of CEHR-BERT. However, these relative performances did not follow the same trend in *hospitalization*, where Bi-LSTM was the second best performing model after CEHR-BERT and outperformed MedBERT and BEHRT. Finally, R-BERT was performing consistently worse than CHER-BERT, suggesting that pre-training played an important role in improving the downstream prediction tasks.

5.2.2. Performance of few-shot learning

One of the advantages of using pre-trained models is the ability to leverage prior knowledge captured from millions of training examples during the pre-training phase. As a result, BERT could be fine-tuned for downstream tasks using a small training set to achieve decent performance. The setup for few-shot learning was similar to the 4-fold evaluation except that we used a subset of the training data in each fold. Specifically, we predefined a list of training percentages (5%, 10%, 20%, 40%, and 80%) with respect to 75% of the training data in each fold in the experiment, and then iterated through the percentages to conduct a separate 4-fold evaluations, in which a subset of the training data was randomly selected based on the training percentage. Apart from randomly sampling the training data, the other configurations including disease phenotypes, baseline models, the process for train/val/test split, and reporting metrics were identical to the disease prediction tasks. Figure 7 shows that CEHR-BERT is the best performing model in terms of AUC and PR-AUC at different training percentages for *t2dm hf*. Following CEHR-BERT, MedBert and BEHRT were other top performers compared to LR, XGB and Bi-LSTM. CEHR-BERT outperformed the second best model MedBert by the same margin throughout the course of this few-show learning experiment. Noticeably, CEHR-BERT fine-tuned with 5% of the training data could achieve an AUC of nearly 0.78 and the PR-AUC of almost 0.29, whereas other models only achieved the AUCs between 0.60 and 0.76 and PR-AUCs between 0.12 and 0.26 despite using up to 80% of the training data. Furthermore, the same trend can be observed for other prediction tasks as well, shown in Figures 8, 9 and 10. The only exception is in *hospitalization*, where CEHR-BERT and LSTM were the first and second best performing models; whereas, other BERTs marginally improved the performance compared to frequency based models.

5.3. Ablation studies

To better understand the contributions of time tokens, time/age embeddings and VTP, we conducted a number of ablation studies. We pre-trained several variations of BERT by excluding one component at a time, then conducted the evaluation described in Disease prediction, and reported results for each variation in Table 4. The best value is highlighted in

Table 3: Average AUC and PR-AUC values and standard deviations for three baseline models and three BERT based models across 4 prediction tasks

		LR	XGB	LSTM	R-BERT	BEHRT	MedBert	CEHR-BERT
t2dm hf	PR	24.8±0.8%	24.8±0.5%	25.8±0.7%	28.1±1.5%	27.1±1.7%	27.37±0.6%	**32.3±1.0%**
	AUC	76.7±0.3%	76.5±0.5%	77.4±0.5%	78.0±0.9%	77.5±0.5%	78.19±0.1%	**80.7±0.6%**
hf readmit	PR	36.3±0.7%	37.6±1.3%	33.3±0.5%	37.1±0.6%	37.4±0.9%	38.0±0.5%	**38.6±0.1%**
	AUC	64.2±0.7%	64.0±0.3%	61.7±0.2%	65.0±0.5%	65.1±0.4%	65.8±0.2%	**66.3±0.2%**
discharge home death	PR	46.7±0.7%	48.5±0.8%	49.1±1.8%	46.7±1.2%	50.7±0.5%	51.4±0.5%	**52.7±0.4%**
	AUC	93.4±0.1%	93.5±0.3%	93.8±0.2%	93.6±0.2%	94.0±0.1%	94.2±0.1%	**94.6±0.1%**
hospitalization	PR	26.9±0.5%	29.0±0.3%	30.0±0.4%	28.0±0.5%	29.4±0.2%	29.5±0.3%	**31.1±0.4%**
	AUC	72.9±0.1%	74.0±0.1%	75.1±0.2%	74.4±0.3%	74.7±0.1%	74.6±0.1%	**75.9±0.1%**

bold in each row. We discuss the contribution of each component separately in the following sections. We also provided the size of each BERT network and the pre-training time in Table 9 so the practitioners could choose the appropriate model based on the use-case.

5.3.1. Assessment of time tokens

To understand the effectiveness of time tokens, we compared our patient representation described in Data processing and patient representation to the existing ones adopted by BEHRT and MedBert. MedBert used a patient representation that only contained the medical concepts and did not include any other artificial tokens. BEHRT used a variation of the MedBert patient representation, where the **SEP** token was inserted between neighboring visits. Figure 5 shows an example of these different EHR representations for a patient's medical history. To perform a fair comparison, we applied the same architecture (described in Methods) to pre-train using these two patient representations, and then we tested the model performances following the same evaluation procedure described in Disease prediction. We will refer to these two BERTs as B-BERT (trained using the BEHRT representation) and M-BERT (trained using the MedBert representation) to distinguish between them. Table 4 shows that CEHR-BERT improved the performances in all tasks as compared to the other patient representations except that M-BERT performed slightly better in terms of PR-AUC in *discharge home death*. Our results indicate that embedding ATTs can effectively capture the temporal information of structured EHR data. One plausible explanation is that ATTs may be treated like any other tokens such that when BERT utilizes them in the self-attention mechanism, those tokens function like a place holder for preserving temporal information, which is then propagated through to the last encoder. In addition, we trained another CEHR-BERT variation where we removed *VS* and *VE* tokens and only kept the ATT tokens, (which we will be refer to as V-BERT). The performance of V-BERT dropped compared to CEHR-BERT, and was comparable to that of B-BERT (*SEP* token inserted between visits), suggesting that *VS* and *VE* play an important role in explicitly defining the boundary of a visit, so that ATT tokens would just function like a regular *SEP* token without their presence.

5.3.2. Assessment of visit type prediction

VTP was designed in this study as a substitute for the original second learning objective NSP to leverage the unique characteristics of structured EHR data. To understand its contribution, we excluded VTP and pre-trained BERT, and then followed the standard evaluation. This BERT will be referred to as NS-BERT. The comparison between NS-BERT and CEHR-BERT in Table 4 demonstrates the improved performances across all prediction tasks attributed to VTP with the exception of *hospitalization*, where NS-BERT slightly outperformed CEHR-BERT but both remained largely similar. Such minor differences may be due to random variation of 4-fold evaluation. The underlying assumption of VTP was that concepts associated with different visit types follow different distributions. Therefore, incorporating VTP into BERT could help learn the representation of concepts. To better understand this, we reported the top 10 most frequent concepts stratified by visit type and domain in Table 11 in Additional Figures and Analyses, which shows that the concept distributions are indeed distinct between visit types. For example, the most

Table 4: Average AUC and PR-AUC values and standard deviations for ablation studies

		M-BERT	B-BERT	NS-BERT	NT-BERT	ALT-BERT	V-BERT	CEHR-BERT
t2dm hf	PR	29.9±1.0%	30.6±0.5%	31.8±1.3%	28.2±0.2%	29.3±0.5%	28.6±0.8%	**32.3±1.0%**
	AUC	79.2±0.3%	79.5±0.5%	80.2±0.6%	78.5±0.4%	76.7±0.2%	78.6±0.1%	**80.7±0.6%**
hf readmit	PR	34.1±0.6%	37.4±0.9%	38.3±0.6%	**39.3±0.7%**	33.3±0.8%	38.6±0.4%	38.6±0.1%
	AUC	62.6±0.3%	65.1±0.2%	65.8±0.1%	**66.4±0.3%**	61.6±0.7%	65.9±0.2%	66.3±0.2%
discharge	PR	**53.1±0.5%**	52.0±1.0%	52.0±0.7%	52.5±0.8%	31.6±3.0%	52.6±1.3%	52.7±0.4%
home death	AUC	94.4±0.3%	94.4±0.1%	94.3±0.1%	94.2±0.3%	87.3±0.9%	94.4±0.1%	**94.6±0.1%**
hospitalization	PR	30.0±0.7%	30.4±0.5%	**31.3±0.7%**	30.8±0.6%	23.4±0.5%	30.6±0.4%	31.1±0.4%
	AUC	74.9±0.3%	75.3±0.3%	**76.1±0.2%**	75.2±0.2%	69.2±0.4%	75.3±0.3%	75.9±0.1%

[1] M-BERT: CEHR-BERT trained on MedBert patient representation
[2] B-BERT: CEHR-BERT trained on BEHRT patient representation
[3] NS-BERT: CEHR-BERT without the second learning objective
[4] NT-BERT: CEHR-BERT without the time/age embeddings
[5] ALT-BERT: modified CEHR-BERT where concept, time and age embeddings are summed together
[6] V-BERT: modified CEHR-BERT where *VS* and *VE* are removed from the patient sequence

frequent concepts associated with inpatient visits relate to childbirth or severe conditions that require immediate hospitalization, while the top outpatient concepts normally pertain to chronic conditions such as hypertension and type 2 diabetes that may not require immediate medical attention but long-term management. Currently, the threshold for masking visit tokens is set to 50%, but we plan to investigate different thresholds to optimize the performance.

5.3.3. Assessment of time/age embeddings

At the input layer, concept, time, and age embeddings are concatenated together, and fed into a FC layer to form temporal concept embeddings, which is then used as the input for CEHR-BERT. To understand the impact of this transformation, we pre-trained a BERT named ALT-BERT, where we summed time embeddings, age embeddings and concept embeddings to form temporal concept embeddings instead of using a FC layer. Table 4 shows that the performance of ALT-BERT was lower compared to that of CEHR-BERT, suggesting that our temporal concept embeddings transformation is more effective than the summation of time/age/concept embeddings. This confirms existing empirical evidence that summing is more rigid than feeding a concatenated product into another the FC layer in terms of fitting the model to the training data.

Furthermore, we wanted to know the impact of using time/age embeddings. Therefore we pre-trained another BERT named NT-BERT, where we disabled all components related to time/age embeddings and only used the concept embeddings. We added the positional encoder (used by Rasmy et al. (2021); Li et al. (2020)) to give the model a basic sense of temporality. Table 4 shows that without time/age embeddings, NT-BERT did perform slightly better in *hf readmit* than CEHR-BERT; however, CEHR-BERT outperformed NT-BERT in all other tasks. In particular, PR-AUC for NT-BERT in *t2dm hf* dropped by 4% compared to CEHR-BERT (from 32.3±1.0% to 28.2±0.2%). Therefore, the benefit of performing the transformation for generating temporal concept embeddings likely outweighs the cost.

6. Discussion

6.1. Understanding artificial time tokens

To understand the functional role of ATTs, we extracted the base embeddings of ATTs and computed 2D features using PCA for visualization. Figure 6 shows that ATTs were arranged from right to left in increasing order of time intervals, where the rightmost and leftmost tokens represent the shortest and longest time intervals respectively. Specifically, the ATT week tokens (including W_0, W_1, and W_2) formed a cluster on the right bottom, and month tokens (in the form of M_n e.g. M_1) seemed to present a linear relationship. In addition, the VS and VE tokens were located far away from time tokens because they represent the start and the end of a visit rather than time intervals, corroborating that their semantics are fundamentally different from time tokens. The analysis supported our assumption that BERT can derive se-

mantics of ATTs and show meaningful relationships between them, thereby learning the hidden knowledge that different time tokens are associated with different groups of concepts.

6.2. Inflated patient sequence

Due to the use of ATTs, we artificially inflated the length of patient sequences, and as a result involuntarily cut-off some records from patient sequences. Whereas, in other sequence models (M-BERT, and Bi-LSTM) we used patient sequences as-is. We calculated the median length (defined as the median number of concepts) as well as 95% sequence length (defined as the number of concepts at 95% percentile) of the patient sequences for different BERT representations across 4 disease prediction tasks, which is shown in Table 10. Among all tasks, patient sequences for *t2dm hf* and *hospitalization* patients have the shortest patient sequences; whereas, *hf readmission* has the richest medical history. Among all patient representations, CEHR-BERT has the longest sequence due to the use of ATTs, and B-BERT has the second longest sequence due to the use of **SEP**. This raised the question of whether cutting off early records in patient sequences could have contributed to the performance boost in evaluations. To address this concern, we re-ran the *hf readmission* analysis using a shorter observation window of 180 days for all models except CEHR-BERT. The motivation was to simulate a scenario where early records were omitted and only the latest records were included. The performance of this modified evaluation was reported in Table 8. It showed that model performances did not improve but dropped when compared to the original evaluation in Table 3 and cutting off the early records in patient sequences did not improve the performance. This comparison suggests the performance gain by CEHR-BERT is attributed to the use of **ATT** tokens.

7. Future work

In this work, we used a context window of 300 for pre-training and conducting experiments. Although we can increase the context window to the BERT sequence limit of 512, patients may have more clinical concepts in their medical histories than the sequence length supported by BERT. To address this limitation, we plan to employ a sliding window strategy, in which we can apply BERT to extract the contextualized embedding representations for each region scanned by the sliding window. Then the regional representations could be combined via a 1D convolution layer or LSTM to generate the final patient representation. Although VTP seems to have improved the results by a small margin, it is not clear whether the improvement could be attributed to random variation, so we need to investigate VTP in the follow-up analysis. Furthermore, we want to include labs to the patient sequence as labs offer an extensive amount of useful information. However, embedding the lab data would require a different strategy as labs are continuous features unlike the discrete data types included in this study. Finally, we will investigate whether or not our methods of incorporating time could be generalized for other models as well.

8. Conclusion

To the best of our knowledge, this is the first study that focuses on incorporating time into BERT for use on structured EHR data, including multiple elements for representing temporal information and leveraging multiple domains of clinical care. CEHR-BERT outperforms existing state of the art BERT-based approaches across a number of different prediction tasks. Based on our results, incorporating time tokens into the patient sequence and combining time/age embeddings with concept embeddings seem to synergistically boost the performance more than any individual modification alone. Therefore, including both into the final CEHR-BERT architecture seems to be the most effective way of capturing the temporal information of structured EHR data. This study was developed using the OMOP common data model, and as a result can be expanded to run on the OHDSI network in order to replicate these results beyond a single database.

Acknowledgments

This work was sponsored by the National Center for Advancing Translational Sciences grant 1U01TR002062 and the Director's Office of National Institutes of Health grants 5U2COD023196 and 3UG3OD023183.

References

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Grin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing*, 2020. doi: 10.1142/9789811215636_0027.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports 2018 8:1*, 8(1):1–12, apr 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-24271-9. URL https://www.nature.com/articles/s41598-018-24271-9.

Youngduck Choi, Chill Yi-I Chiu, and David Sontag. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41, 2016. URL /pmc/articles/PMC5001761//pmc/articles/PMC5001761/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001761/.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019. ISBN 9781950737130.

Benjamin S. Glicksberg, Riccardo Miotto, Kipp W. Johnson, Khader Shameer, Li Li, Rong Chen, and Joel T. Dudley. Automated disease cohort selection using word embeddings from Electronic Health Records. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 23(212669):145, 2018. URL /pmc/articles/PMC5788312//pmc/articles/PMC5788312/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5788312/.

Sara Bersche Golas, Takuma Shibahara, Stephen Agboola, Hiroko Otaki, Jumpei Sato, Tatsuya Nakae, Toru Hisamitsu, Go Kojima, Jennifer Felsted, Sujay Kakarmath, Joseph Kvedar, and Kamal Jethwani. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC Medical Informatics and Decision Making*, 18(1), jun 2018. doi: 10.1186/S12911-018-0620-Z. URL /pmc/articles/PMC6013959//pmc/articles/PMC6013959/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6013959/.

George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, Martijn J. Schuemie, Marc A. Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R. Rijnbeek, Johan Van Der Lei, Nicole Pratt, G. Niklas Norén, Yu Chuan Li, Paul E. Stang, David Madigan, and Patrick B. Ryan. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. In *Studies in Health Technology and Informatics*, volume 216, pages 574–578. IOS Press, 2015. ISBN 9781614995630. doi: 10.3233/978-1-61499-564-7-574.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2Vec: Learning a Vector Representation of Time. jul 2019. URL https://arxiv.org/abs/1907.05321v1.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, dec 2014. URL https://arxiv.org/abs/1412.6980v9.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for Electronic Health Records. *Scientific Reports 2020 10:1*, 10(1):1–12, apr 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-62922-y. URL https://www.nature.com/articles/s41598-020-62922-y.

Matthew McDermott, Bret Nestor, Evan Kim, Wancong Zhang, Anna Goldenberg, Peter Szolovits, and Marzyeh Ghassemi. A comprehensive EHR timeseries pre-training benchmark. *ACM CHIL 2021 - Proceedings of the 2021 ACM Conference on Health, Inference, and Learning*, pages 257–278, apr 2021. doi: 10.1145/3450439.3451877.

Kenney Ng, Steven R. Steinhubl, Christopher DeFilippi, Sanjoy Dey, and Walter F. Stewart.

Early Detection of Heart Failure Using Electronic Health Records: Practical Implications for Time before Diagnosis, Data Diversity, Data Quantity and Data Density. *Circulation. Cardiovascular quality and outcomes*, 9(6):649, nov 2016. doi: 10.1161/CIRCOUTCOMES.116.002797. URL /pmc/articles/PMC5341145//pmc/articles/PMC5341145/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5341145/.

Xueping Peng, Guodong Long, Tao Shen, Sen Wang, Jing Jiang, and Michael Blumenstein. Temporal Self-Attention Network for Medical Concept Embedding. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2019-November: 498–507, sep 2019. doi: 10.1109/icdm.2019.00060. URL https://arxiv.org/abs/1909.06886v1.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1: 2227–2237, feb 2018. URL https://arxiv.org/abs/1802.05365v2.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine 2021 4:1*, 4(1):1–13, may 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00455-y. URL https://www.nature.com/articles/s41746-021-00455-y.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of Graph Augmented Transformers for Medication Recommendation. *IJCAI International Joint Conference on Artificial Intelligence*, 2019-August:5953–5959, jun 2019. URL https://arxiv.org/abs/1906.00346v2.

Yang Xiang, Jun Xu, Yuqi Si, Zhiheng Li, Laila Rasmy, Yujia Zhou, Firat Tiryaki, Fang Li, Yaoyun Zhang, Yonghui Wu, Xiaoqian Jiang, Wenjin Jim Zheng, Degui Zhi, Cui Tao, and Hua Xu. Time-sensitive clinical concept embeddings learned from large electronic health records. *BMC Medical Informatics and Decision Making*, 2019. ISSN 14726947. doi: 10.1186/s12911-019-0766-3.

Jinghe Zhang, Kamran Kowsari, James H. Harrison, Jennifer M. Lobo, and Laura E. Barnes. Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record. *IEEE Access*, 6:65333–65346, oct 2018. doi: 10.1109/access.2018.2875677. URL https://arxiv.org/abs/1810.04793v3.

Appendix A. Prediction Tasks

A prediction task can be phrased as the following, "among a particular group of people, who will go on and experience some event". One can think of this problem as defining a target cohort that represents the initial group of people, and an outcome cohort that represents the subset of the initial group who will experience a particular event, e.g. among the type 2 diabetes patients, who will go on and develop heart failure. Both target and outcome cohorts can be defined as a group of people who satisfy certain inclusion criteria for a certain period of time. Typically, a cohort definition includes a cohort entry event and a set of inclusion criteria (an exclusion criterion can be thought of as an inclusion criterion with 0 occurrence). Specifically, the cohort entry event defines the index date, at which the patients enter the cohort, and the inclusion criteria add more constraints to the cohort if applicable, such as the requirements of certain diagnosis, medications, procedures or temporal relationships among criteria, and etc. In addition, a prediction window needs to be specified for generating the ground truth labels for the given target and outcome cohorts, if the outcome index date falls between the index date of the target cohort and the prediction window, we will declare the case to be positive, and otherwise negative.

In terms of prediction time range, we use observation window, hold-off window and prediction window to collect data in different time periods. One patient's medical history is built on an event sequence. Each event represents one medical engagement within a visit. It could be condition occurrence, medication exposure, procedure occurrence and measurement etc. In this paper, we only focus on the first three types of events. We extract features using data from the observation window followed by a hold-off window to avoid same target concepts being included into feature construction. The prediction window is right after the index event. And the goal is to predict the occurrence of any outcome event. With the datasets, we could train, test and validate the model. Figure 2 visualizes the cohort definition that is constructed based on all events in patient medical history.

We validate our model by applying it to 4 downstream prediction problems. Table 5 lists the cohort definition parameters for the prediction tasks.

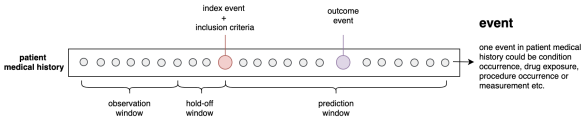

Figure 2: Cohort Definition and Prediction Windows

A.1. Feature Engineering

Frequency based features were used for LR and XGboost, and generated as the following 1) International Classification of Diseases (ICD) ICD-9 or ICD-10 codes were rolled up to 2 or 3 digit codes for procedure and diagnosis depending on where the dot character is in the code e.g. 44.98 and I50.0; 2) for procedure records encoded by Current Procedural Terminology (CPT), they were rolled up to the second level from the root; 3) all medications were rolled up to the ingredient level; 4) for the codes that couldn't be rolled up, the original codes were kept; 5) the frequencies of the rolled-up codes were calculapted within the observation window.

A.2. Type 2 Diabetes Mellitus patients who developed Heart Failure

The target cohort consists of patients who had T2DM (type 2 diabetes mellitus) in the medical history. The index event of T2DM is patient encounters with condition concept ids or medication exposures of anti-diabetes medications. We also exclude any patients with pre-existing T2DM, type 1 diabetes, diabetes insipidus, gestational diabetes, secondary diabetes and neonatal diabetes.

The outcome cohort is a subset of the target cohort who developed heart failure during the prediction window. We defined Heart Failure cohort as patients who were diagnosed with heart failure, at least one lab test with high BNP results, received any treatment including mechanical circulatory support, artificial heart associated procedure, diuretic agent, vasoactive agent or dialysis procedure.

For each of the criteria, we construct a concept set with a group of OMOP concept ids. In Table 6 & Table 7, we list all related concept sets and concept ids from OMOP vocabulary.

Table 5: Cohort Definition Parameters

cohort name	index event	inclusion criteria	outcome event	observation window	hold-off window	prediction window
t2dm hf	Type 2 Diabetes Mellitus condition occurrence or medication exposure	No pre-existing diabetes, type 1 diabetes, diabetes insipidus, gestational diabetes, secondary diabetes, neonatal diabetes	Heart Failure	unbounded	0	unbounded
hf readmit	Heart Failure condition occurrence	At least one HF treatment, lab test or medication exposure	Readmission (In-patient visit)	360 days	0	30 days
discharge home death	Discharge to home	Following an inpatient visit	Death	360 days	0	360 days
hospitalization	EHR start	Number of visits between 2 and 30	In-patient visit	540 days	180 days	720 days

A.3. Heart Failure patients who were readmitted within 30 days

The target cohort contains patients who were admitted into hospital due to heart failure . The index event is an inpatient visit with a heart failure diagnosis (316139). Patients in the target cohort who were readmitted into hospital within 30 days will be in the outcome cohort. The concept_ids of inpatient visits are 9201 and 262. In this case, the prediction window is 30 days.

A.4. Patients who were discharged and died within one year

The target cohort is patients who had an inpatient visit and were discharged to home. The outcome cohort is patients who died within one year after being discharged. The index event is inpatient visit with visit concept_id as 9201 or 262 and discharge_to_concept_id is home or other nursing facilities. The outcome event is death. The prediction window is 360 days.

A.5. Hospitalization

The hospitalization cohort has a special structure showed in Figure 3 than the generalized cohort structure. The index event is when the patient had the first visit in the hospital. And the observation window is post the index event. The outcome event is an inpatient visit. We only include patients who had visit occurrences between 2 to 30 during the observation window to make sure patients had enough data points and also remove any outliers.

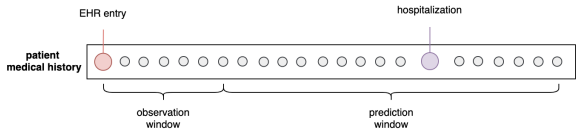

Figure 3: Hospitalization Cohort Definition and Prediction Windows

Appendix B. Additional Figures and Analyses

Table 6: T2DM Cohort Related Concepts

Domain	**Concept Set**	**OMOP concept ids**
Condition	T2DM	443238, 201820, 442793, 4016045
	pre-existing T2DM	40769338,43021173,42539022, 46270562
	type 1 diabetes	201254, 4019513, 40484648
	diabetes insipidus	30968, 438476
	gestational diabetes	4058243
	secondary diabetes	195771
	neonatal diabetes	193323
Medication	Metformin	1503297
	Chlorpropamide	1594973
	Glimepiride	1597756
	Glyburide	1559684
	Glipizide	1560171
	Tolbutamide	1502855
	Tolazamide	1502809
	Pioglitazone	1525215
	Rosiglitazone	1547504
	Sitagliptin	1580747
	Saxagliptin	40166035
	Alogliptin	43013884
	Linagliptin	40239216
	Repaglinide	1516766
	Nateglinide	1502826
	Miglitol	1510202
	Linagliptin	40239216
	Acarbose	1529331
	Insulin	35605670, 35602717, 1516976, 1502905, 46221581, 1550023, 35198096, 42899447, 1544838, 1567198, 35884381, 1531601, 1588986, 1513876, 19013951, 1590165, 1596977, 1586346, 19090204, 1513843, 1513849, 1562586, 19090226, 19090221, 1586369, 19090244, 19090229, 19090247, 19090249, 19090180, 19013926, 19091621, 19090187

Table 7: Heart Failure Cohort Related Concepts

Domain	**Concept Set**	**OMOP concept ids**
Condition	Heart Failure	316139
Medication	Diuretic Agent	4186999, 956874, 942350, 987406, 932745, 1309799, 970250, 992590, 907013
	Vasoactive Agent	1942960
Measurement	High B-type Natriuretic Peptide (BNP) > 500 pg/mL	4307029
	NT-proBNP > 2000 pg/mL	1594973
Procedure	Mechanical Circulatory Support	45888564, 4052536, 4337306, 2107514, 45889695, 2107500, 45887675, 43527920, 2107501, 45890116, 40756954, 4338594, 43527923, 40757060, 2100812
	Artificial Heart Associated Procedure	4144390, 4150347, 4281764, 725038, 725037, 2100816, 2100822, 725039, 2100828, 4337306, 4140024, 4146121, 4060257, 4309033, 4222272, 4243758, 4241906, 4080968, 4224193, 4052537, 4050864

Table 8: Average AUC and PR-AUC values for LR, XGB, LSTM, BEHRT, and MedBert for *hf readmission* using the 180-day observation window

	LR	XGB	LSTM	BEHRT	MedBert
PR-AUC	36.6±0.7%	37.1±0.6%	36.3±0.1%	37.2±0.3%	31.7±0.4%
AUC	64.7±0.3%	64.1±0.7%	64.2±0.4%	65.0±0.3%	59.5±0.1%

Table 9: Trainable and Non trainable parameters counts for M-BERT, B-BERT, NS-BERT, ALT-BERT, V-BERT and CEHR-BERT.

	M-BERT	B-BERT	NS-BERT	ALT-BERT	V-BERT	CEHR-BERT
Trainable Parameters	9083060	9083060	8811380	9044596	9085108	9085364
NonTrainable Parameters	0	0	0	0	0	0
Time to pre-train per epoch	8.5h	8.5h	7.5h	8.5h	8.5h	8.5h

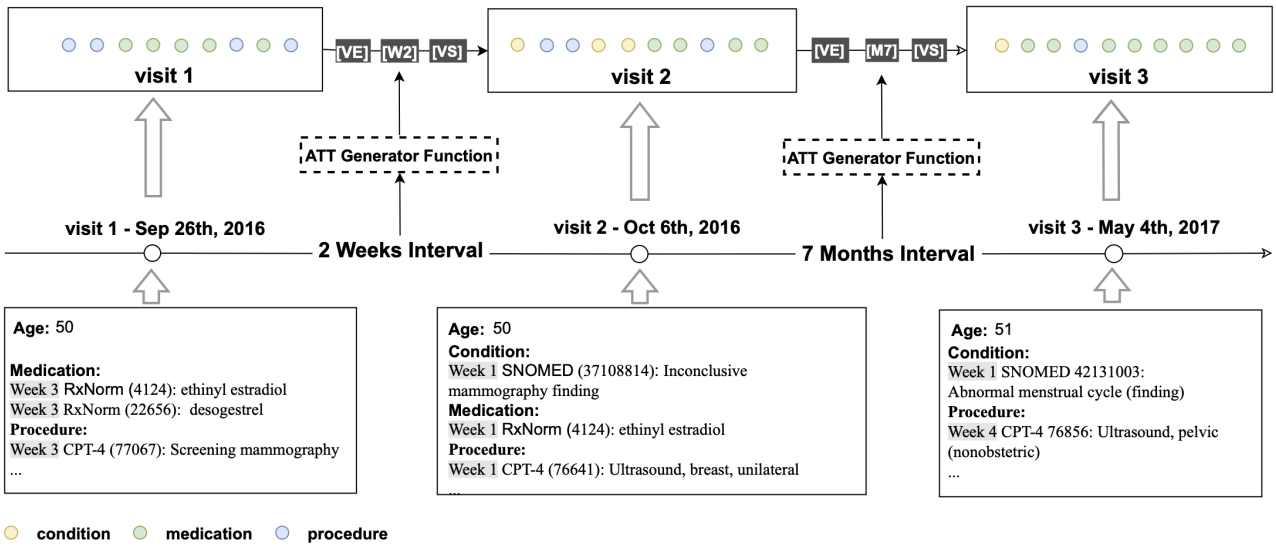

Figure 4: Patient sequence representation and artificial time tokens

Table 10: Patient sequence lengths for 3 BERT representations across 4 disease predictions

		HF readmission	*Discharge home Death*	*T2DM HF*	*Hospitalization*
median length	CEHR-BERT	123	64	21	19
	b-bert	98	47	15	12
	m-bert	86	38	13	9
95% length	CEHR-BERT	608	330	366	108
	b-bert	520	257	246	78
	m-bert	481	223	189	66

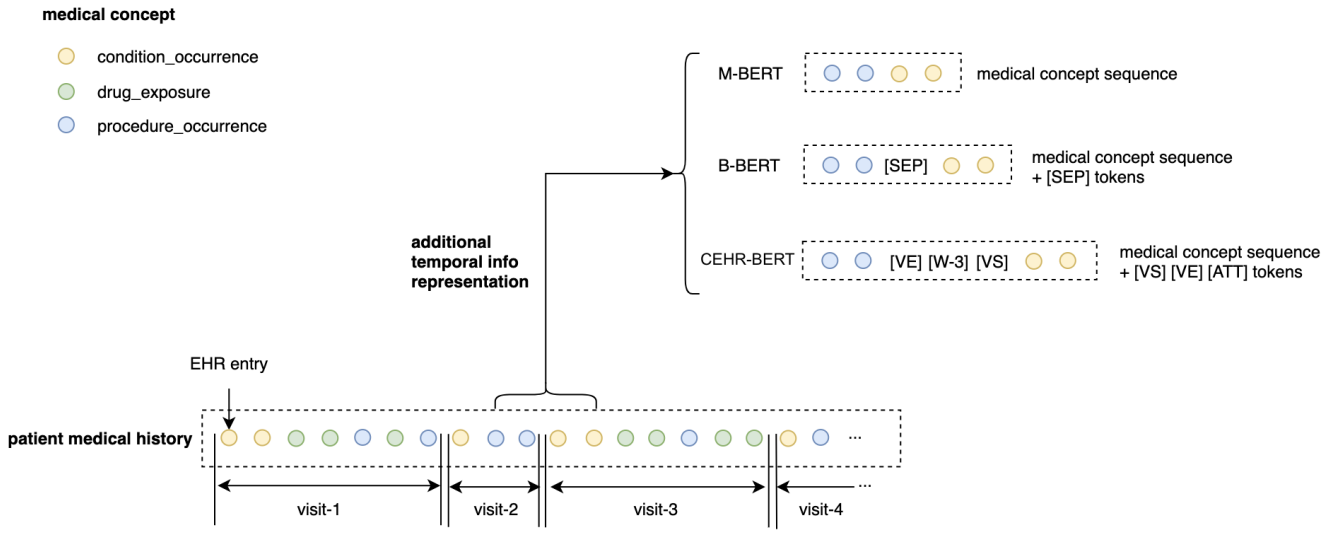

Figure 5: EHR Representation of patient medical history

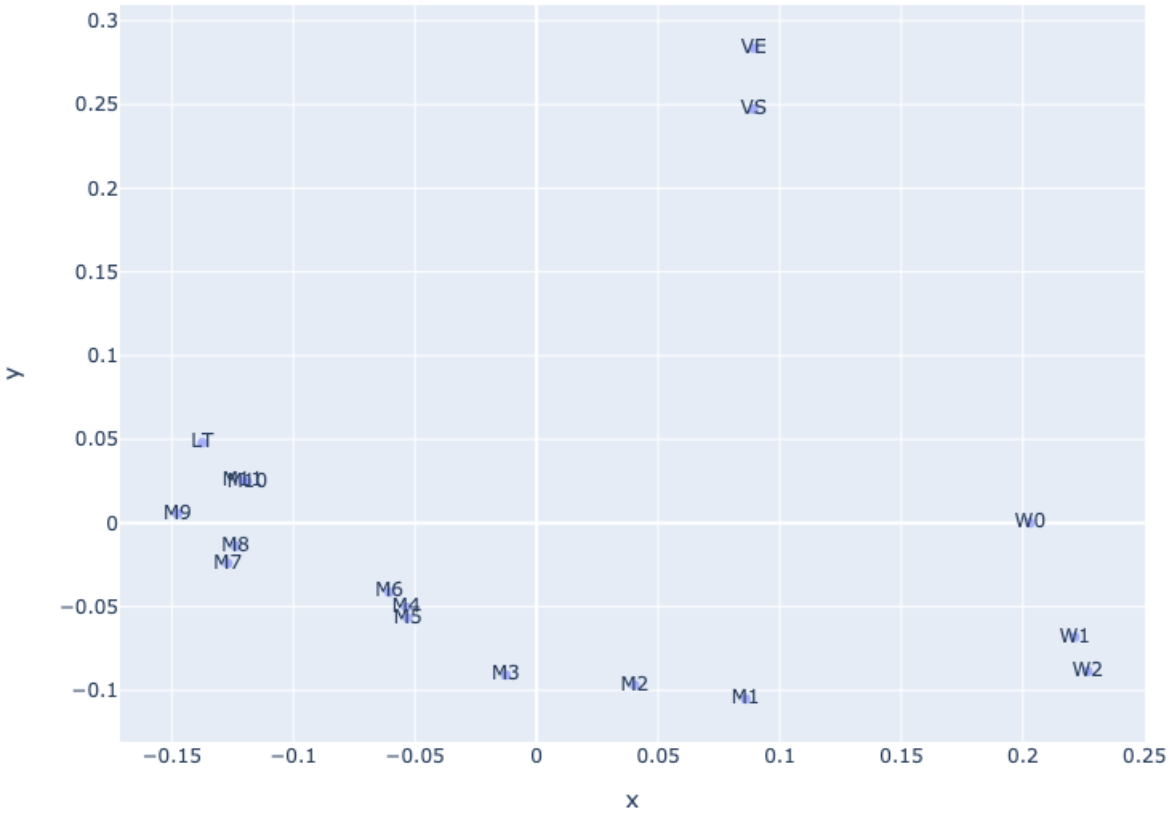

Figure 6: 2d visualization of the Artificial Time Tokens added to CEHR-BERT. The base embeddings of those tokens were extracted from CEHR-BERT, and PCA was run to extract the 2d features for visualization.

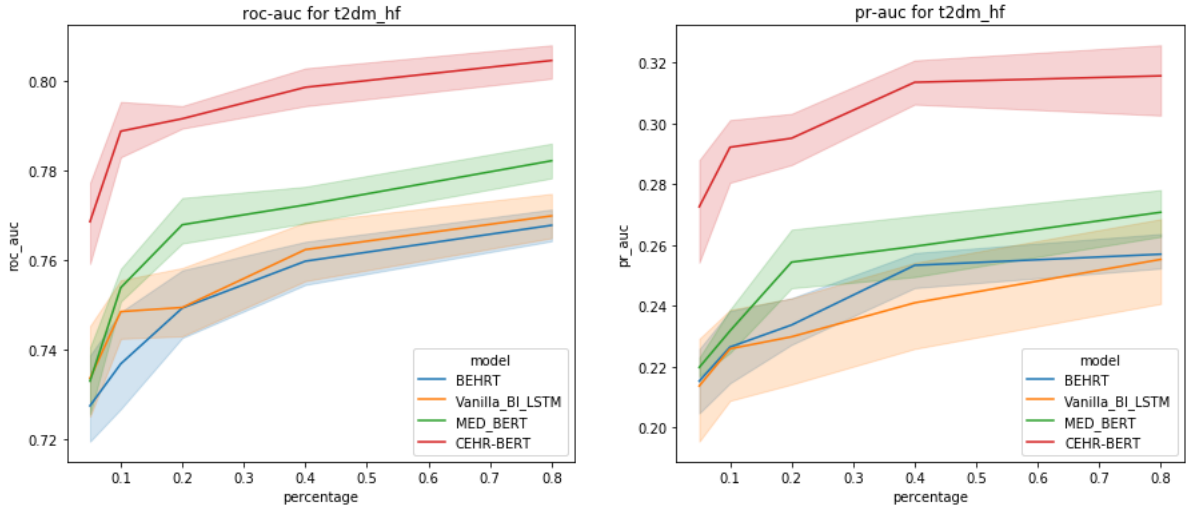

Figure 7: AUC and PR-AUC at different training percentages for all the models for few-shot learning task for *hf readmission* are plotted against

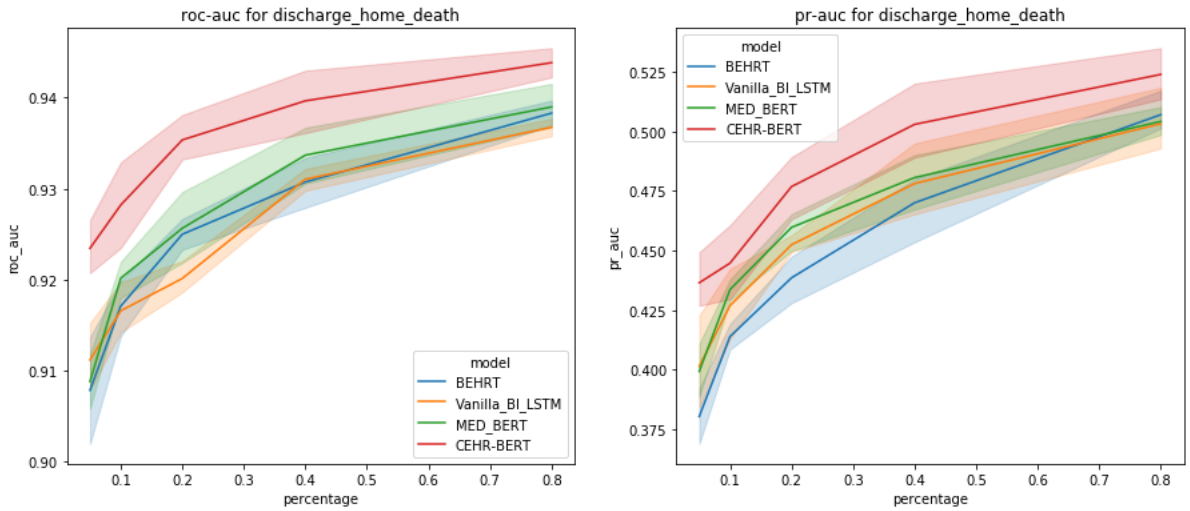

Figure 8: AUC and PR-AUC at different training percentages for all the models for few-shot learning task for *discharge home death* are plotted against

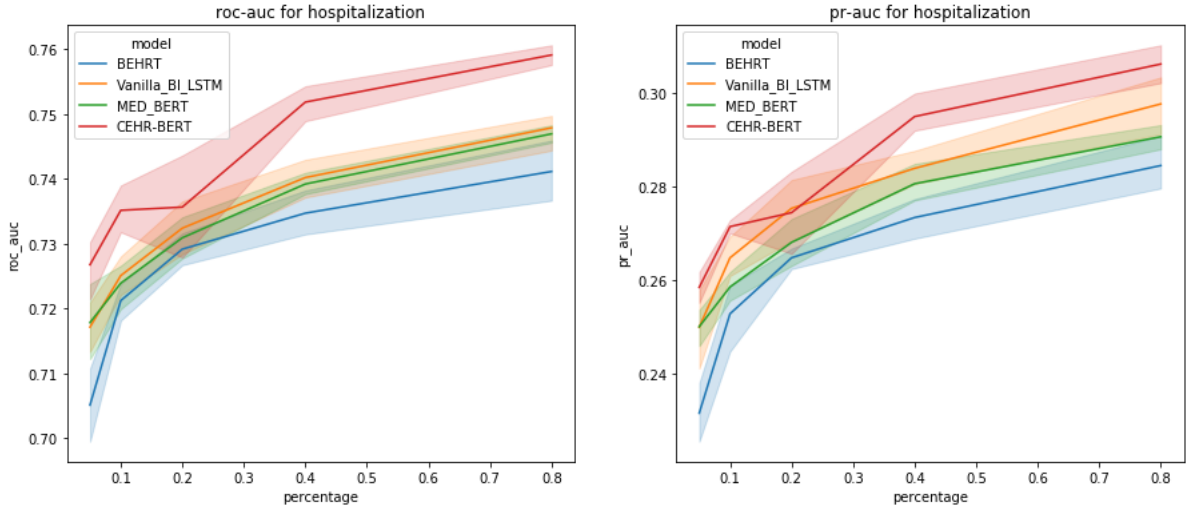

Figure 9: AUC and PR-AUC at different training percentages for all the models for few-shot learning task for *hospitalization* are plotted against

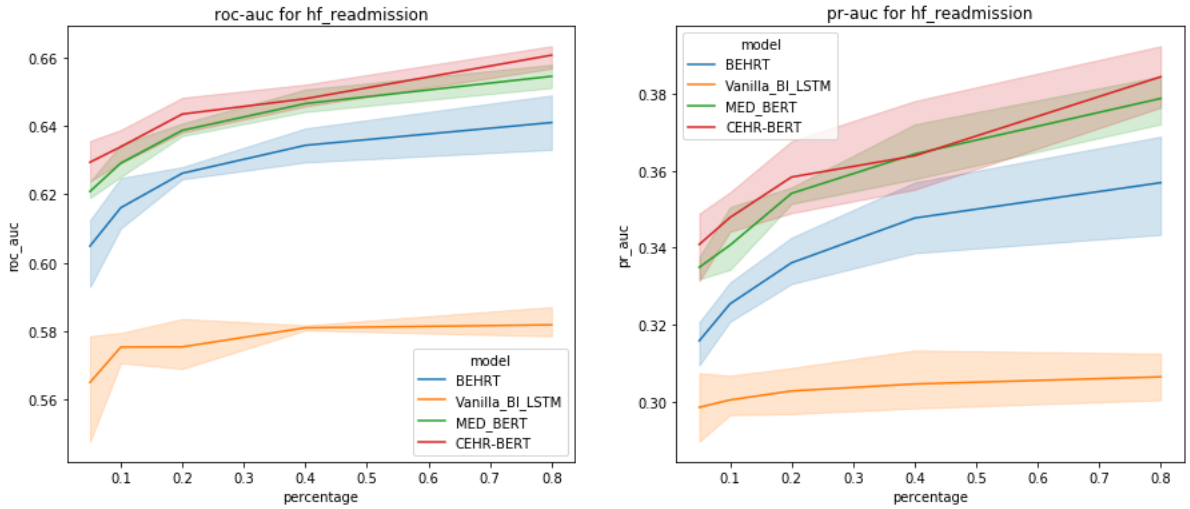

Figure 10: AUC and PR-AUC at different training percentages for all the models for few-shot learning task for *hf readmission* are plotted against

Table 11: Top 10 most frequent condition concepts associated with each visit type

Visit Type	Rank	Condition Concepts	Patient Count	Percentage of the Visit Type Patients
Outpatient Visit	1	Essential hypertension	310794	13.00%
	2	Abdominal pain	172393	7.21%
	3	Chest pain	164499	6.88%
	4	Hyperlipidemia	160068	6.70%
	5	Finding related to pregnancy	140044	5.86%
	6	Joint pain	117134	4.90%
	7	Dyspnea	114580	4.79%
	8	Low back pain	114031	4.77%
	9	Pure hypercholesterolemia	113437	4.75%
	10	Unplanned pregnancy	110584	4.63%
Inpatient Visit	1	Single live birth	279820	35.01%
	2	Essential hypertension	150002	18.77%
	3	Finding related to pregnancy	121545	15.21%
	4	Postpartum finding	100345	12.55%
	5	Delivery normal	66666	8.34%
	6	Late effect of medical and surgical care complication	63224	7.91%
	7	Congestive heart failure	55631	6.96%
	8	Coronary arteriosclerosis	53763	6.73%
	9	Hyperlipidemia	51459	6.44%
	10	Diabetes mellitus without complication	51069	6.39%
Emergency Room and Inpatient Visit	1	Essential hypertension	64914	30.09%
	2	Finding related to pregnancy	43878	20.34%
	3	Single live birth	41606	19.29%
	4	Hyperlipidemia	39051	18.10%
	5	Acute renal failure syndrome	34763	16.11%
	6	Postpartum finding	30382	14.08%
	7	Anemia	24841	11.51%
	8	Urinary tract infectious disease	23831	11.05%
	9	Chest pain	22768	10.55%
	10	Dehydration	22402	10.38%
Home Visit	1	Essential hypertension	144	45.71%
	2	Malaise	97	30.79%
	3	Constipation	89	28.25%
	4	Major depression, single episode	61	19.37%
	5	Cough	52	16.51%
	6	Dementia	43	13.65%
	7	Dementia with behavioral disturbance	42	13.33%
	8	Disorder due to infection	42	13.33%
	9	Slow transit constipation	41	13.02%
	10	Hyperlipidemia	40	12.70%

Visit Type	Rank	Condition Concepts	Patient Count	Percentage of the Visit Type Patients
Office Visit	1	Essential hypertension	62118	5.83%
	2	Hyperlipidemia	36152	5.76%
	3	Gastroesophageal reflux disease without esophagitis	26766	5.61%
	4	Vitamin D deficiency	26442	4.65%
	5	Cough	25727	4.48%
	6	Pure hypercholesterolemia	21321	4.38%
	7	Obesity	20555	4.23%
	8	Chronic pain	20091	4.07%
	9	Dyspnea	19387	4.01%
	10	Fatigue	18656	3.73%
Health examination	1	Chronic pain	369	16.83%
	2	Shoulder joint pain	248	11.31%
	3	Musculoskeletal finding	172	7.84%
	4	Low back pain	158	7.20%
	5	Interstitial lung disease	126	5.75%
	6	Pain in right knee	124	5.65%
	7	Pain in left knee	110	5.02%
	8	Postoperative state	103	4.70%
	9	Difficulty walking	93	4.24%
	10	Lumbago with sciatica	89	4.06%
Emergency Room Visit	1	Abdominal pain	153538	13.77%
	2	Essential hypertension	123530	11.08%
	3	Chest pain	99274	8.91%
	4	Viral disease	95402	8.56%
	5	Headache	94671	8.49%
	6	Fever	92601	8.31%
	7	Cough	89758	8.05%
	8	Acute upper respiratory infection	88679	7.96%
	9	Acute pharyngitis	74356	6.67%
	10	Asthma	73301	6.58%