

# Learning the Joint Representation of Heterogeneous Temporal Events for Clinical Endpoint Prediction

Luchen Liu<sup>1</sup>, Jianhao Shen<sup>1</sup>, Ming Zhang<sup>1\*</sup>, Zichang Wang<sup>1</sup>, Jian Tang<sup>2,3\*</sup>

<sup>1</sup>School of EECS, Peking University, Beijing China

<sup>2</sup>HEC Montreal, <sup>3</sup>Montreal Institute for Learning Algorithms

liuluchen@pku.edu.cn, jhshen@pku.edu.cn, mzhong\_cs@pku.edu.cn, dywzc123@163.com, tangjianpku@gmail.com

## Abstract

The availability of a large amount of electronic health records (EHR) provides huge opportunities to improve health care service by mining these data. One important application is clinical endpoint prediction, which aims to predict whether a disease, a symptom or an abnormal lab test will happen in the future according to patients' history records. This paper develops deep learning techniques for clinical endpoint prediction, which are effective in many practical applications. However, the problem is very challenging since patients' history records contain multiple heterogeneous temporal events such as lab tests, diagnosis, and drug administrations. The visiting patterns of different types of events vary significantly, and there exist complex nonlinear relationships between different events. In this paper, we propose a novel model for learning the joint representation of heterogeneous temporal events. The model adds a new gate to control the visiting rates of different events which effectively models the irregular patterns of different events and their nonlinear correlations. Experiment results with real-world clinical data on the tasks of predicting death and abnormal lab tests prove the effectiveness of our proposed approach over competitive baselines.

## Introduction

The volume of electronic health records (EHR) is expanding at a staggering rate, providing a great opportunity for machine learning and data mining researchers to analyze these data so as to provide better health care service. An important application of machine learning in health care is predicting the clinical endpoints such as a disease, symptom, or laboratory abnormality based on patients' historical records.

This paper develops effective deep learning techniques for clinical endpoint prediction since deep learning techniques have been proved effective for predictive analysis in a variety of applications such as image recognition (?), speech recognition (?), and natural language understanding (?). The goal of deep learning is to learn effective semantic representations of the high-dimensional data such as images, speeches and natural language. Therefore, our goal is to effectively represent patients' historical records.

However, the problem is challenging since patients' historical records contain a variety of heterogeneous temporal

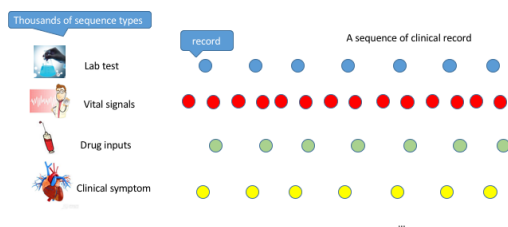


Figure 1: Heterogeneous Temporal Events. The sampling rates of different events vary significantly from each other. Different kinds of events are highly correlated.

events such as different lab tests, routine vital signals, diagnosis, and drug administrations (See Fig. 1 as an example). The visiting rates of different events vary significantly. For example, a patient may take a blood test every morning while take a temperature test every two hours. Besides, there is a high level of dependency among different kinds of events. For instance, some diagnosis are made according to the results of some lab tests. As a result, these heterogeneous temporal events yield heterogeneous event sequences consisting of thousands of correlated event types, the visiting rate of which varies significantly.

In the literature, learning representations of sequences are widely studied especially in the domain of speech recognition and natural language understanding. The state-of-the-art approaches for sequence modeling are recurrent neural networks (RNNs) with the Long Short-term Memory (LSTM) units. RNNs are commonly used for modeling homogeneous sequences, but it is nontrivial to apply them for modeling heterogeneous event sequences. There are some recent works based on multi-task Gauss Process (MTGP) for modeling the correlations between multiple sequences. However, the computational cost of MTGP is too expensive for EHR data since there are thousands of types of events. Therefore, we are seeking an approach that is able to: (1) effectively model the irregular visiting patterns of different events; (2) model the complex nonlinear relationships between different events; (3) scale up a large number of different types of events.

In this paper, we propose such an approach called Heterogeneous Event LSTM (HE-LSTM) for learning the joint rep-

\*corresponding authors

representation of heterogeneous event sequences. Our approach is an extension of Phased LSTM (?), which was recently proposed and is used to model irregular event-based sequential data. Compared to the vanilla LSTM model, Phased LSTM (?) adds a new time gate, which is able to naturally integrate inputs from several sensors of arbitrary sampling rates. But Phased LSTM is not suitable for modeling the heterogeneous event sequence with thousands of event types in longitudinal EHR data. Our proposed model extends it by modeling correlated heterogeneous events with multi-scale sampling rates. Each event type and its attributes are embedded and fed into HE-LSTM. The HE-LSTM is equipped with an event gate controlled by the event type embeddings and their timestamps. With the help of the event gates, the HE-LSTM can perfectly trace the temporal information of different event types in the long heterogeneous event sequence by asynchronously sample important and related events in the heterogeneous event sequence. Therefore, the representation of heterogeneous temporal events can be updated based on the dependency of the current input event and other events maintained in the HE-LSTM.

We conduct extensive experiments on real-world clinical data. Experiment results on the tasks of death prediction and abnormal lab test prediction prove that our proposed approach outperforms competitive baselines. Our proposed approach can be widely used in modeling data collected from sensors with arbitrary sampling rates, such as data collected from mobile sensors.

Our main contributions are:

- We formulate the clinical endpoint prediction task based on EHR data as a representation learning problem of heterogeneous temporal events consists of asynchronous clinical records from multiple sources.
- We propose a novel model called HE-LSTM for learning the representations of heterogeneous event sequence. The model effectively models the multi-scale sampling rates of different kinds of events and their temporal dependency.
- We conducted experiment on real-world clinical data on the tasks of death and abnormal lab tests. Promising results prove the effectiveness of our proposed approach over competitive baselines.

## Related Works

### Clinical Endpoint Prediction

There are plenty of works trying to solve the clinical endpoint prediction problem. However, many of these works only use a small subset of the whole EHR sequences in order to avoid dealing with the high-dimensional event types. Some works select a subset of the clinical events from the EHR data according to the expertise of physicians (?). For instance, Alaa only uses a set of 21 (temporal) physiological streams comprising a set of 11 vital signs and 10 lab test scores to predict ICU admission (?). Some techniques select 50 time series from the whole set of EHR data, and transformed the fixed-size subset into a new latent space using the hyper-parameters of multi-task GP(MTGP) models. They then calculate the similarity of patient’s records in the new hyper-parameter space (?). It is notable that manually

selecting only a fraction of clinical sequences from original EHR data as the input brings out expert bias, thus these works seldom make full use of the important information of original data.

Most works ignore the content or value of clinical events, and only use the type information of clinical events to predict the endpoints (?). Specifically, some approaches train the semantic embeddings for different categories of clinical events for endpoint predictions (?). RETAIN uses two reversed recursive neural networks(RNN) generating attention variables of sequential ICD-9 code groups for the prediction tasks (?). There are some works using convolution neural network(CNN) to model irregular medical codes for future risk predictions (?). These works only exploit the type information of historical clinical events to make predictions, ignoring the fine-grained varying attributes of the events. Our work is to address the issue by utilizing the rich type information of clinical events as well as the content and values of the events.

## Deep Learning Models for Sequential Data

Standard RNNs trained with stochastic gradient descent have difficulty learning long-term dependencies (i.e. spanning more than 10 time steps) encoded in the input sequences owing to the vanishing gradient (?). The problem has been addressed for example by using a specialized neuron structure in Long Short-Term Memory (LSTM) networks (?) that maintains constant backward flow in the error signal.

In the Clockwork RNN (CW-RNN) (?), the hidden layer is partitioned into separate modules, each processing inputs at its own temporal granularity, making computations only at its prescribed clock rate. In this way, the fixed clock periods help to contain long-term dependencies.

Phased LSTM (?) is a state-of-the-art RNN architecture for modeling event-based sequential data. It extends LSTM by adding the time gate. The gate has three phases: it rises from 0 to 1 in the first phase and drops from 1 to 0 in the second phase, which are active states. During the third phase, the model is in the inactive state. Updates to  $c_t$  and  $h_t$  are permitted only in the active state. The Phased LSTM network can achieve fast convergence in most experiments, owing to the fact that the auto-sampling on the long sequential data conducted by the time gate maintains derivative error in the longer back propagation.

However, these models only focus on learning long-term dependencies in homogeneous sequences, lacking the ability to capture the various and complex temporal dependencies in heterogeneous temporal events, which usually exist in EHR data.

## Task Definition

Here are some notations and the definition of the task.

**Heterogeneous Events Sequence** The heterogeneous event is defined as the triple  $e_i = (type, value, time)$ . *type* is the category of event, *value* is the attribute of the event, *type* and *value* of  $e_i$  are logged at *time*. It is noteworthy

that the attributes *values* of different event types can be either numerical or categorical variable. For example, the attributes of a lab test, e.g. lactate blood test, is numerical while the attribute of the clinical status, e.g. ectopia type, is categorical variable (i.e. fusion beats, nodal bigeminy).

Heterogeneous events are merged in the ascending order of the record time into a triple sequence  $\{e_i\}$ . We denote the heterogeneous event sequence in a period of time  $[T_{start}, T_{end}]$  as  $\{e_i\}_{T_{start} \leq e_i.t \leq T_{end}}$ .

**Clinical Endpoint Prediction Task** The clinical endpoint prediction task is formulated as follow: given a clinical heterogeneous event sequence  $\{e_i\}_{T_{start} \leq e_i.t \leq T_{end}}$ , and a binary label  $\hat{y}$  for the target endpoint occurring at  $T_{end} + 24$  hours, the objective is to predict what the target endpoint is in 24 hours using  $\{e_i\}_{T_{start} \leq e_i.t \leq T_{end}}$ .

In this paper, we aim to dynamically predict two endpoint outcomes base on the heterogeneous event sequence of patient data in EHR. In the first “death prediction dataset”, the endpoint outcome is death in either hospital or discharge to home. In the second “lab test result prediction dataset”, the endpoint outcome is either an abnormal result of the potassium lab test, or clinical stability.

## Proposed Method

In this section, we introduce the technical details about our proposed model. The overall view of our model is illustrated in Figure 2.

### Event type embedding and attribute encoding

To help the HE-LSTM to trace temporal information of various kinds of events, we use “event type embedding” and “attribute encoding” to embed the type and attributes of the high dimensional events into compact continuous vectors, which can be trained end-to-end with the following HE-LSTM.

An event  $e_i = (type, value, time)$  of the sequence will be embedded into three parts to feed the HE-LSTM for the endpoint prediction. The three input including embedding vector of event type  $s$ , the event attribute encoding vector  $x$  and the scale variable time  $t$ .

The event type vector  $s$  carries the information of the event category of  $e_i$ , and is constructed only by the one hot representation *type* of event type. Similar to word embedding (?), it will provide a low-dimension vector of the event type with semantic meaning in clinical field. The embedding lookup matrix  $C_{type} \in \mathbb{R}^{N \times M}$ , where  $N$  is the embedding dimension and  $M$  is the number of event types, is established for further training. The event type vector  $s$  is given by:

$$s = C_{type} \times type \quad (1)$$

The event attribute encoding vector  $x$  represents the combining information of both event type *type* and the attribute of the event *value*, which is the main input of the following HE-LSTM. Each event has two kinds of attributes. One is categorical with the one hot representation  $value_c \in \{0, 1\}^C$ , where  $C$  is the number of values of

categorical attributes of all the event types. The other is numerical with the one hot representation  $value_n \in \mathbb{R}^U$ , where  $U$  is the number of all numerical attribute types of all the event types. Notice that the event attribute vector  $value = [value_c, value_n]$ .

Each value of categorical attributes is assigned with a vector from  $V_c \in \mathbb{R}^{N \times C}$ , where  $N$  is the embedding dimension. As for numerical attributes, they are associated with a value encoding vector in  $V_n \in \mathbb{R}^{N \times U}$ , where  $N$  is the embedding dimension.

The representing vector of a record  $x$  is mainly decided by its event type  $s$ , however the event attributes also carries lots of information for modeling patients. The different values of the same event type, such as the abnormal label in a lab test event, can lead to distinct estimates for the patient’s future health status. The other important part of  $x$  is a disturbance from the numerical attribute values. For instance, the high numerical value of the lactate blood lab test event indicates potential health problem of the patient, while the low value does not offer much information. Finally, to combine the three parts of information, the attribute encoding vector  $x$  is given by:

$$x = s + V_c \times value_c + \tanh(V_n \times value_n) \quad (2)$$

where  $V_c$  and  $V_n$  are parameters to learn.

### Heterogeneous Event LSTM

Long short-term memory (LSTM) units (?) (Fig. 2(a)) is an important ingredient of modern deep RNN architectures. We first define their update equations in a commonly-used version in the following:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + w_{ic} \circ c_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + w_{fc} \circ c_{t-1} + b_f) \quad (4)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \quad (5)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + w_{oc} \circ c_{t-1} + b_o) \quad (6)$$

$$h_t = o_t \circ \tanh(c_t) \quad (7)$$

The main difference from classical RNNs is the use of the gating functions  $i_t$ ,  $f_t$ ,  $o_t$ , which represent the input, forget, and output gate at time  $t$  respectively.  $c_t$  is the cell activation vector, whereas  $x_t$  and  $h_t$  represent the input feature vector and the hidden output vector respectively. The gates use the typical sigmoid function  $\sigma$  and tanh nonlinear function  $\tanh$  with weight parameters  $W_{ih}$ ,  $W_{fh}$ ,  $W_{oh}$ ,  $W_{ix}$ ,  $W_{fx}$ , and  $W_{ox}$ , which connect the different inputs and gates with the memory cells and outputs, as well as biases  $b_i$ ,  $b_f$ , and  $b_o$ . The cell state  $c_t$  itself is updated with a fraction of the previous cell state that is controlled by  $f_t$ , and a new input state created from the element-wise product, denoted by  $\circ$ , of  $i_t$  and the output of the cell state nonlinearity  $\tanh$ . Optional peephole (?) connection weights  $w_{ic}$ ,  $w_{fc}$ ,  $w_{oc}$  further influence the operation of the input, forget, and output gates.

HE-LSTM extends the LSTM model by adding a new event gate  $j_{s,t}$ . The event gate has two factors — an event

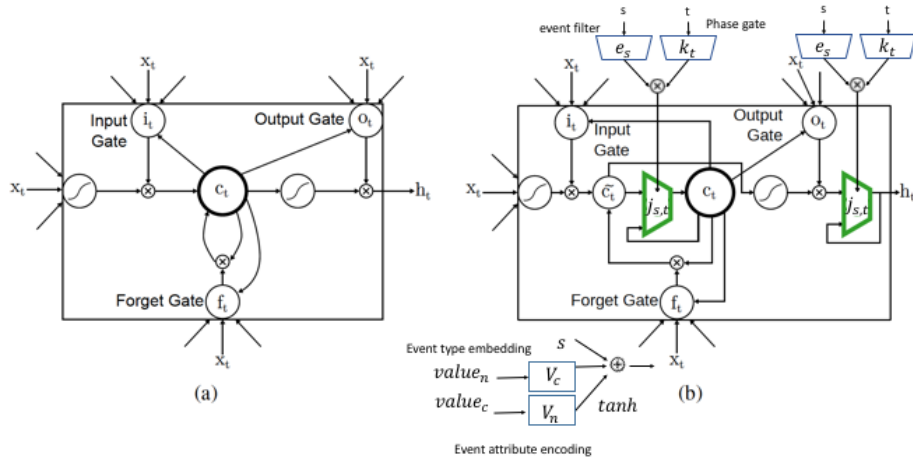


Figure 2: Model architecture. (a) Standard LSTM model. (b) HE-LSTM model, with event gate  $j_t$  consist of the event filter  $e_s$  and phase gate  $k_t$  separately controlled by the event type  $s$  and timestamp  $t$ . In the HE-LSTM formulation, each neural in the cell value  $c_t$  and the hidden output  $h_t$  can be updated during an “open” phase by only some certain types of events; otherwise, the previous values are maintained.

filter and a phase gate. The event filter only allows the information of a certain cluster of events to fuse into the corresponding memory cell, so that each cell will only trace a particular group of events. Collaborated with the phase gate, the event filter can help the network to maintain the temporal information of the different events in multi-scaled sampling rates. The dependency of the heterogeneous events will be easier to capture by the diverse and long memory of correlated events.

The opening and closing of this event gate is controlled by the event type embedding  $s$  and an independent rhythmic oscillation specified by the phase gate (?) with three parameters. And updates to the cell state  $c_t$  and  $h_t$  are permitted only when the gate is opened.

One factor of the event gate, the event filter  $e_s$ , for each neuron is a feed forward network with a hidden layer of size  $L$  with  $\tanh$  activation function as following.

$$e_s = \sigma(W_{em} \tanh(W_{ms}s + b_m) + b_e) \quad (8)$$

where  $W_{em} \in \mathbb{R}^{H \times L}$ ,  $W_{ms} \in \mathbb{R}^{L \times N}$ ,  $b_e \in \mathbb{R}^H$  and  $b_m \in \mathbb{R}^L$  are parameters to learn.

Considering the multi-scale sampling rates of the events, we extend the event filter  $e_s$  with a time factor proposed in phased LSTM (?) by three parameters:  $t$ ,  $r_{on}$  and  $s$ , where  $t$  represents the real-time period of the gate,  $s$  represents the phase shift and  $r_{on}$  is the ratio of the open phase to the full period.  $t$ ,  $r_{on}$  and  $s$  are learned by training. Therefore,  $j_{s,t}$  is formally defined as:

$$\phi_t = \frac{(t - s) \bmod \tau}{\tau}, k_t = \begin{cases} \frac{2\phi_t}{r_{on}}, & \text{if } \phi_t < \frac{1}{2}r_{on} \\ 2 - \frac{2\phi_t}{r_{on}}, & \text{if } \frac{1}{2}r_{on} < \phi_t < r_{on} \\ \alpha\phi_t, & \text{otherwise} \end{cases} \quad (9)$$

$$j_{s,t} = e_s \odot k_t \quad (10)$$

Different from traditional RNNs for single sequential data and even sparser variants of RNNs (?), updates in HE-LSTM can optionally be performed at irregularly sampled time points  $t$  for different event types. This allows the RNNs to learn the multi-scale rhythm of related events and work with asynchronously sampled heterogeneous temporal event data. We use the shorthand notation  $c_l = c_{t_l}$  for cell states at time  $t_l$  (analogously for other gates and units), and let  $c_{l-1}$  denote the state at the previous update time  $t_{l-1}$ . We can then rewrite the regular LSTM cell update equations for  $c_l$  and  $h_l$  (from Eq. 5 and Eq. 7), using proposed cell updates  $\tilde{c}_l$  and  $\tilde{h}_l$  mediated by the event gate  $j_l$ :

$$\tilde{c}_l = f_l \odot c_{l-1} + i_l \odot \tanh(W_{cx}x_l + W_{ch}h_{l-1} + b_c) \quad (11)$$

$$c_l = j_l \odot \tilde{c}_l + (1 - j_l) \odot c_{l-1} \quad (12)$$

$$\tilde{h}_l = o_l \odot \tanh(\tilde{c}_l) \quad (13)$$

$$h_l = j_l \odot \tilde{h}_l + (1 - j_l) \odot h_{l-1} \quad (14)$$

The HE-LSTM formulation ensures the flexible allocation and retain of information of each event clusters. Each neuron of the memory cell and hidden layer of HE-LSTM states can be updated only during the open periods of the event gate. In other words, only the information of a certain cluster of events' records can flow into this certain neuron in its own phase. This is because the event filter  $e_s$ , one of the factor of the event gate  $j_{s,t}$ , can be seen as a binary classifier to chose the cluster of event types responsible for each neuron. Besides, the neuron maintains a perfect memory during its closed phase, i.e.  $c_l = c_{l-\delta}$  if  $j_{l'} = 0$  for  $t_l \leq l' \leq t_{l-\delta}$ . Thus, other neurons, tracing other events can directly use the information of this cluster of events even they are far away

from each other in term of the sequence index. Because of this allocation mechanism, HE-LSTM can have much diverse and longer memory for modeling the dependency of multiple events.

We use a sigmoid layer to predict the true label  $\hat{y}_t$  of the learned representation vector of sequence in the given decision times.

$$y_t = \text{sigmoid}(\mathbf{w}_p \mathbf{h}_t + b_p) \quad (15)$$

where  $\mathbf{w}_p \in \mathbb{R}^N$  and  $b_p$  are parameters to learn.

We use cross-entropy to calculate the classification loss of the prediction  $y_t$  and true label  $\hat{y}_t$  of each sample as follows:

$$\text{Loss}(\hat{y}_t, y_t) = \frac{1}{N} \sum_{1 \leq t \leq N} (\hat{y}_t \times \ln y_t + (1 - \hat{y}_t) \times \ln(1 - y_t)) \quad (16)$$

We can sum up the losses of all the samples in one mini-batch to get the total loss for back propagation.

## Experiments

The source code of MIMIC-III EHR data preprocessing and the proposed model is available and can be found at <https://github.com/pkusjh/HELSTM>.

### Data Description and Experimental Settings

We set up two data sets for evaluation of the models from one real clinical data source. MIMIC-III (Medical Information Mart for Intensive Care III) is a large, freely-available database comprising de-identified health-related data relating to over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.

We extract all kinds of events from the MIMIC-III database to get the initial event type set(18192 kinds of events in total).The statistics of the event types with top frequency are listed in Table.1. By merging the heterogeneous events into triple sequence, we get a set of clinical event sequences. We drop out the sparse event types, whose frequency in total is less than 2500.

We extract episodes of patients, which are 24 hours before the occurrence time of each endpoint, from these event sequences as samples. And the upper bound of the record number of the samples is 1000. All the resulting sample events are labeled according to the target endpoint outcome in each task.

The statistics of the final clinical multiple sequences in two datasets are summarized in Table 2.

Each dataset is split into 3 parts with fixed proportions, namely training set(70%), validation set(10%) and evaluation set(20%). The data in validation set is used to select hyper-parameters of the proposed and comparing models and to conduct ‘‘early stop’’ while training, in which the samples may be different because of cross-validation. The evaluation set, the details of which is non-transparent for us in the process of training and parameter selection, will then be only used to calculate and report the evaluation metrics for comparison.

event sources	e.g. event types	# of types
lab test	HEMATOCRIT, WHITE BLOOD CELLS	525
vital signal	Heart Rate, Respiratory Rate	385
drug input	0.9% Normal Saline, Dextrose 5%	60
clinical symptom	Ectopy Type Motor Response	2382
procedure	electrocardiogram Invasive Ventilation	19
clinical output	Urine gastric retentive oral dosage	17

Table 1: Statistics of the event type

Dataset	# of samples	# of events	Avg timespan
death	24301(8%)	20290879	3d 15h 58m
lab test	784583(11%)	41006177	192d 22h 45m

Table 2: Statistics of the datasets(the percentage in the second column is the positive sample rate)

For all the results presented in this section, the networks, implemented with Theano (?),were trained with Adam (?) set to default learning rate parameters.

### Comparing Methods

We compare HE-LSTM to the following methods.

- **Independent LSTM** We use LSTM to model each homogeneous event independently and average the resulting representations into a logistic regression layer. Because the computational cost of thousands of independent LSTM exceed our tolerance, we select 25 important events as it was done in many works (?).

- **Independent LSTM (shared weight)** This model is the same as the previous one, except that the weights in each single LSTM is shared and all events are used as the input of the model.

- **RETAIN RETAIN** (?) mimics physician practice by modeling the EHR data in a reverse time order, and a two-level RNN generating attention variables of sequential data can provides interpretation of the prediction.

- **LSTM + event embedding & attr encoding** We use the event embedding in the first part of proposed method section as the input of traditional LSTM. Logistic regression is applied to the top hidden layer.

- **Clock-work RNN** Clockwork RNN (?) described in related works section.

- **Phased LSTM** Phased LSTM (?) described in related works section.

### Evaluating Metrics

All methods listed above can produce predict scores instead of binary labels, and the data for target prediction tasks are imbalanced labeled. So metrics for binary labels such as accuracy are not suitable for measuring the performance. Similar to the work (?; ?), we adopt the area under ROC curves

Methods	death		lab test	
	AUC	AP	AUC	AP
Independent LSTM	$0.8771 \pm 0.0005$	$0.5573 \pm 0.0006$	$0.7196 \pm 0.0006$	$0.2969 \pm 0.0008$
Independent LSTM(shared weight)	$0.8064 \pm 0.0005$	$0.5301 \pm 0.0006$	$0.5308 \pm 0.0005$	$0.1098 \pm 0.0005$
Phased LSTM	$0.8474 \pm 0.0005$	$0.4900 \pm 0.0075$	$0.7722 \pm 0.0007$	$0.3575 \pm 0.0026$
Clock-work RNN	$0.8400 \pm 0.0001$	$0.7181 \pm 0.0003$	$0.6516 \pm 0.0002$	$0.2208 \pm 0.0003$
RETAIN	$0.8967 \pm 0.0011$	$0.5808 \pm 0.0114$	$0.7325 \pm 0.0022$	$0.3096 \pm 0.0052$
LSTM + event embedding & attr encoding	$0.9466 \pm 0.0002$	$0.7445 \pm 0.0007$	$0.7231 \pm 0.0028$	$0.3021 \pm 0.0014$
HE-LSTM	<b><math>0.9516 \pm 0.0003</math></b>	<b><math>0.7687 \pm 0.0011</math></b>	<b><math>0.7987 \pm 0.0008</math></b>	<b><math>0.3914 \pm 0.0013</math></b>

Table 3: Performance of clinical endpoint prediction tasks

(Receiver Operating Characteristic curves) and area under PRC (Precision-Recall curves) for evaluation. Both reflect the overall quality of predicted scores at each decision time, according to their true labels.

- **the Area under ROC Curve(AUC)** of comparing  $y_i$  with the true label  $\hat{y}_i$ . AUC is robust to imbalanced positive/negative prediction labels, making it appropriate for evaluating the classification accuracy in the endpoint prediction tasks.

- **Average Precision(AP)** Average precision (?) emphasizes ranking positive samples higher. It is the average of precisions computed at the point of each positive samples in the ranked sequence in ascending order of predict score:

$$AP = \frac{\sum_{r=1}^N (P(r) \times I(r))}{\text{number of positive samples}} \quad (17)$$

$$P(r) = \frac{|\{\text{positive samples of rank } r \text{ or less}\}|}{r} \quad (18)$$

where  $r$  is the rank,  $N$  the total number of samples,  $I()$  a index function on the positive sample of a given rank, and  $P(r)$  precision at a given cut-off rank.

This metric is also referred to geometrically as the area under the Precision-Recall curve.

- **Cross Entropy** that measures the model loss on the test set. The loss can be calculated by Eq (16).

## Quantitative Results

Table.3 shows the area under ROC and AP of different methods on death and lab test datasets respectively. From the results in Table.3, we draw the following conclusions:

Firstly, models considering the dependency of correlated events types outperform all the independent sequential models and the proposed HE-LSTM achieves the best performance. For example, on death prediction task, RETAIN, LSTM and HE-LSTM improve the AP of lab test prediction by around 4.3%, 2.4% and 32.1% respectively compared to the best of “independent LSTM” model without weight share of the parameters in each independent LSTMs. The similar results have been shown in other experiments and metrics. Furthermore, our model achieves the highest performance among these heterogeneous sequential models. For example, on lab test prediction task, HE-LSTM improves the AP by 26.2% and 29.4% compared to RETAIN and LSTM. Besides, the improvements on AUC are 9.0% and 10.4% respectively. We can draw the conclusion that the dependency information of correlated clinical temporal events is useful

in endpoint prediction tasks and learning joint representations is more effective to model the temporal dependency of different events of EHR data compared to simple independent sequential models.

Secondly, compared to the densely updating recurrent neural networks, the RNNs adaptive to the sampling rate pattern of events make more improvement of the prediction performance. For example, clock-work RNN improve the AP of death prediction by 29.0% and 33.9% compared to the two kinds of independent LSTMs, while the improvements of AUC and AP are 7.0% and 20.7% for phased LSTM compared to the best of independent LSTMs in the lab test prediction task. We can draw the conclusion that multi-scaled sampling rate pattern of events is effective for endpoint prediction, which makes the model concentrate more on the important events in different phases other than treating all clinical events equally in the long sequence.

Thirdly, HE-LSTM achieves the best performance on all datasets and all evaluation metrics. HE-LSTM outperforms all sparsely updating recurrent neural networks and heterogeneous sequential models on each metrics of two datasets. Models solely utilizing multi-scale sampling pattern in event sequence or models straight-forwardly merging different type of events are not the best choice for clinical endpoint prediction in EHR data. Take the result of death prediction for example, HE-LSTM improves the AUC and AP by 12.4% and 7.0% respectively compared to the best of sparsely updating methods without the event type embedding and attribute encoding modules. The improvements of HE-LSTM compared to the heterogeneous sequential models without event gates are 3.4% and 30.3% in average in term of AUC and AP. We can draw the conclusion that the proposed HE-LSTM effectively improve the performance because of the joint effects of tracing the temporal dependency of heterogeneous events and adaptively fitting their multi-scaled sampling patterns.

**Experiment on variations of the event gate** To evaluate the effect of the components in the event gate  $j_{s,t}$ , we replace  $j_{s,t}$  in (Eq 10) with its factors, namely phase gate  $k_t$  and event filter  $e_s$  while remaining the other parts of the model identical. The results on two datasets are list in table 4, including AUC, AP and cross entropy on test data as well as the values of the three metrics when the first training epoch is finished.

The event filter mainly helps to improve the performance of clinical endpoint prediction tasks by modeling the depen-



Methods		Phase gate	Event filter	Event gate
death prediction	AUC(1st epoch)	0.9301	0.9105	<b>0.9370</b>
	AUC	0.9471	<b>0.9518</b>	0.9516
	AP(1st epoch)	0.6856	0.6048	<b>0.7094</b>
	AP	0.7467	0.7679	<b>0.7687</b>
	Entropy(1st epoch)	0.1561	0.1835	<b>0.1479</b>
abnormal lab test prediction	Entropy	0.1369	0.1301	<b>0.1297</b>
	AUC(1st epoch)	0.7050	0.6747	<b>0.7275</b>
	AUC	0.7945	0.7559	<b>0.7987</b>
	AP(1st epoch)	0.2752	0.2403	<b>0.2965</b>
	AP	0.3875	0.3410	<b>0.3914</b>
	Entropy(1st epoch)	0.3373	0.3448	<b>0.3298</b>
	Entropy	0.3019	0.3178	<b>0.3003</b>

Table 4: Performance with different settings of the event gate

dependency of heterogeneous events. Both the event gate and the event filter achieve good performance in all metrics and both datasets when the training is finished. For example, the event gate and the event filter improve the AUC of death prediction by 0.5% and 0.5% compared to the phase gate, while the improvements of AP are 2.8% and 2.9% and the improvements of entropy are 4.9% and 5.2%.

The phase gate helps to achieve a fast convergence in the early stage of training by fitting the multi-scaled sampling rates of different events. HE-LSTM and the model with only phase gate get much higher performance in all metrics and both datasets in the first epoch of training. Take results in lab test task for example, the phase gate and the event gate improve the AUC in first epoch by 4.6% and 7.9% compared to the event filter, while the improvements of AP are 14.5% and 23.3% and the improvements of entropy are 2.2% and 4.3%.

From these comparisons, we draw the conclusion that the event filter and the phase gate collaborates jointly in modeling the dependency in heterogeneous temporal events with the multi-scale sampling rates, which leads to the accurate and efficient performance on the clinical endpoint prediction task.

**Experiment on varying length of multiple sequential data** To evaluate the ability to model the temporal dependency of heterogeneous temporal events of our proposed architect and the other baselines, we feed the trained models multiple events in test set with various length, in the range of 20 to 1000, as input. From figure 3, we can draw the following conclusions:

Firstly, temporal information is effective for endpoint prediction tasks. The performances of most models improve with the increase of the input sequence length. Especially, the performance increases sharply when the length of input sequence is less than 200.

Secondly, HE-LSTM is better at handling the dependency of heterogeneous temporal events than other models. When the input sequence is short, the performances of different models are similar. The reason lies in the fact that, for short sequence input, the combination of independent representations of a single event makes less difference from the joint representation of heterogeneous events in HE-LSTM. But when the input sequence get longer and longer, the performance of our model steadily increase from 0.7551 to

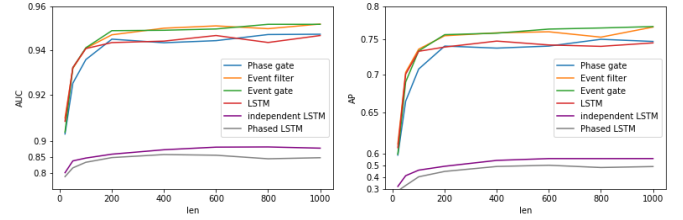


Figure 3: the performance on death prediction task with varying input length of the heterogeneous event sequence data

0.7687 in term of AP and from 0.9482 to 0.9516 in term of AUC. The performance of other models remained almost unchanged at almost 0.9465 of AUC and 0.7434 of AP.

**Different initial period** To explore the effect of the event filter in the event gate when modeling heterogeneous sequential EHR data, we compare the performance of the proposed HE-LSTM with the reduced HE-LSTM, of which the event filter factor in the event gate is removed. We use different initial periods of  $\tau$  during training for death prediction task. The period was drawn uniformly in the exponential domain, comparing four sampling intervals  $\exp(U(1, 2))$ ,  $\exp(U(2, 3))$ ,  $\exp(U(3, 4))$ , and  $\exp(U(4, 5))$  for each model. The results in Figure.4 show that the initialization of  $\tau$  affects the performance of both models. But HE-LSTM is more robust to the initialization. For example, the improvements of HE-LSTM compared to the one without event filter are 4.1%, 4.1%, 2.8% and 6.6% on average. We can draw the conclusion that, with the help of event filter, the event gate can be more adaptive to multi-scale sampling rates of the events in the heterogeneous temporal sequence.

## Conclusion

In this paper, we propose a novel HE-LSTM model to learn joint representations of heterogeneous temporal events for clinical endpoint prediction. Our model can adaptively fit the multi-scaled sampling rates of events in the heterogeneous event sequence. By tracing the temporal information of different kinds of events in the long sequence, the temporal dependency of different types of events can be cap-

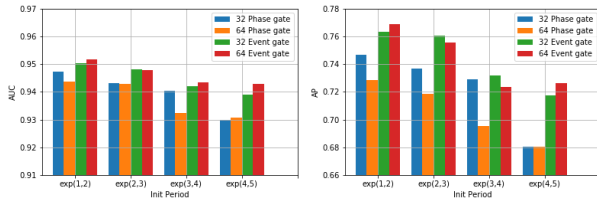


Figure 4: Different initial period

tured in our learned representations. Experimental results with real-world clinical data on the tasks of predicting death and abnormal lab tests prove the effectiveness of our proposed approach over competitive baselines.

### Acknowledgement

This paper is partially supported by the National Natural Science Foundation of China (NSFC Grant Nos.91646202, 61772039 and 61472006). Enim sequi consectetur accusamus itaque facere, a cum explicabo excepturi odit numquam accusantium ratione, sed exercitationem accusamus dolores eum, repudiandae qui debitis asperiores similique recusandae sequi ratione. Facilis ipsa hic, doloreque odio beatae quo quae nostrum asperiores veritatis, pariatur nihil debitis possimus provident nisi quod porro iure molestiae aut, quisquam nihil corrupti laborum sed amet aliquam incidunt iste obcaecati nesciunt labore, harum natus quas esse voluptates vel praesentium eos voluptas labore ad quidem? Architecto rem odio sunt doloreque earum repellendus temporibus quaerat exercitationem eum ducimus, inventore est repellendus ratione rem pariatur, accusamus eligendi aut ratione explicabo sunt possimus id doloreque impedit, consequatur provident earum, pariatur nihil tempore nostrum impedit inventore est ex debitis. Dignissimos totam optio, sunt iure ullam animi doloreque illum nam beatae placeat illo architecto vero, eos explicabo odio dolorum quia et ducimus, eius officia cupiditate ipsam distinctio sint earum, quae minima deleniti voluptatem molestiae tenetur magnam doloribus iste excepturi? Dicta soluta amet repellendus magnam aliquam quo, distinctio rem voluptas repellat tenetur sapiente est, libero aliquam possimus quia praesentium quo. Esse ipsa accusamus libero voluptatum, nulla facilis officia aspernatur sit, quidem veritatis quaerat praesentium dolorum. Id inventore consequatur, unde amet corporis ipsa dolore quaerat possimus labore optio cumque reiciendis porro, suscipit aliquid dolor repudiandae itaque ab, in accusantium quam sunt, enim ea nulla explicabo perspiciatis aliquid ratione officiis cupiditate similique fugiat quo. Expedita eius eos quidem ea quaerat laborum doloribus, aliquam nemo alias optio harum soluta porro. Deleniti amet ratione quae, totam quod magni error vel delectus animi blanditiis neque, debitis eveniet sed magni dolor excepturi odit voluptas dolore necessitatibus sit, aspernatur debitis quam molestiae autem perspiciatis error reprehenderit, id dicta quae dolorum laboriosam rem amet asperiores earum natus labore numquam? Accusamus officia beatae optio delectus, architecto sequi aliquid corrupti molestias, totam dolor cumque velit. Quis sit dolor eum

vel fugiat voluptatum vero, aut dolore esse officiis iste error id itaque tempora expedita? Soluta at quia hic inventore repellat fugiat aliquid libero sit, ducimus veniam deserunt harum cumque odit architecto accusantium labore sapiente enim, eum fugiat est, eos aliquid optio fugiat voluptate, doloreque accusamus velit doloribus ut consectetur quo itaque voluptate similique pariatur? Libero beatae labore, vel fugit sed aperiam laborum animi corporis culpa sint maxime minima? Dolorem quos illum, dicta aperiam quaerat inventore facere minima, nesciunt ea ab officia similique praesentium distinctio molestias sit reiciendis earum, at eligendi culpa voluptates porro ut distinctio cum impedit adipisci cumque dolore. Architecto sed pariatur harum commodi officiis porro qui, minus rerum soluta praesentium dicta, minima omnis praesentium minus enim quidem, aliquid inventore ad eveniet corrupti at et animi quidem pariatur, totam fuga quas aliquid saepe similique dicta iure veniam aut optio nemo? Quod cum obcaecati eaque esse deserunt velit accusamus quo sequi error, ipsa accusamus aspernatur nobis asperiores quo excepturi nam amet totam, aliquid maxime eveniet cupiditate, quo mollitia praesentium tempore perspiciatis illum beatae officia ex quidem ipsam? Odit nemo quod quia, deserunt ratione aliquam ea id dolore magni voluptates saepe, in maxime necessitatibus omnis atque labore dolor. Pariatur soluta perspiciatis ipsa non iusto aut recusandae accusamus et, adipisci similique porro minus tenetur illo, temporibus assumenda ex a repellat deserunt adipisci voluptas debitis architecto, alias unde omnis architecto magni incidunt quasi? Aut cum vel quidem vero numquam adipisci amet et, laudantium reprehenderit sapiente accusantium libero voluptatem fugit harum tempore, obcaecati corrupti minima quia itaque quae molestiae, aliquam iure eius minima dicta. Nemo atque doloreque, voluptatem odit ratione laborum, laudantium voluptas nihil nulla ad mollitia, quibusdam excepturi temporibus ipsam recusandae voluptatibus harum necessitatibus quasi quidem unde distinctio. Recusandae beatae necessitatibus harum maxime illum adipisci natus autem animi, consequuntur eos tempore aliquid laborum distinctio veritatis, at eos asperiores temporibus pariatur, a rem vitae beatae ipsa eum optio in nostrum repudiandae adipisci, fugiat tenetur at quae ut perspiciatis quod quas voluptas. Quaerat nobis placeat perferendis illum quidem itaque officiis reprehenderit, aspernatur error inventore eum blanditiis eaque sint fugiat est aliquam? Doloremque amet magni, ad atque similique odit cumque modi tenetur, aperiam perferendis sunt dolore, illum ab praesentium mollitia inventore deleniti, mollitia omnis natus voluptates hic aliquid cupiditate rerum. Blanditiis veniam illum culpa rerum doloribus perspiciatis hic, facere eum illum. Incidunt minima animi, unde iure dignissimos doloribus earum nisi quia mollitia suscipit, quos cum modi porro nostrum facilis qui delectus aliquam ex adipisci, magni quod libero voluptas officiis sequi aliquam adipisci ab provident. Ipsam voluptatem quo dolorum, perferendis ab quam non quidem earum, similique laboriosam ex sed, eveniet atque facilis, quas placeat aperiam modi aliquam nihil tempore? Impedit perspiciatis unde blanditiis maxime, cum a debitis quis ipsum, illum quae tempore. Dolor voluptates sit sint voluptas accusamus, facere