BMC Medical Informatics and Decision Making

TECHNICAL ADVANCE

Open Access

# Interpretable time-aware and co-occurrence-aware network for medical prediction

Chenxi Sun[1,2*], Hongna Dui[3] and Hongyan Li[1,2]

## Abstract

**Background:** Disease prediction based on electronic health records (EHRs) is essential for personalized healthcare. But it's hard due to the special data structure and the interpretability requirement of methods. The structure of EHR is hierarchical: each patient has a sequence of admissions, and each admission has some co-occurrence diagnoses. However, the existing methods only partially model these characteristics and lack the interpretation for non-specialists.

**Methods:** This work proposes a time-aware and co-occurrence-aware deep learning network (TCoN), which is not only suitable for EHR data structure but also interpretable: the co-occurrence-aware self-attention (CS-attention) mechanism and time-aware gated recurrent unit (T-GRU) can model multilevel relations; the interpretation path and the diagnosis graph can make the result interpretable.

**Results:** The method is tested on a real-world dataset for mortality prediction, readmission prediction, disease prediction, and next diagnoses prediction. Experimental results show that TCoN is better than baselines with 2.01% higher accuracy. Meanwhile, the method can give the interpretation of causal relationships and the diagnosis graph of each patient.

**Conclusions:** This work proposes a novel model—TCoN. It is an interpretable and effective deep learning method, that can model the hierarchical medical structure and predict medical events. The experiments show that it outperforms all state-of-the-art methods. Future work can apply the graph embedding technology based on more knowledge data such as doctor notes.

**Keywords:** Medical prediction, Interpretable deep learning, Electronic health records, Disease correlation

## Background

Electronic Health Records (EHRs) are increasingly popular and widely used in hospitals for better healthcare management. A typical EHR dataset consists of much patient information, including demographic information and medical information. The medical information is an irregular hierarchical patient-visit-code (patient-admission-diagnosis) form, shown in Fig. 1a: (1) Each patient has many visit records as he/she may go to see a doctor many times. The visit records have corresponding time stamps and form a sequence; (2) Each visit contains many codes, which are usually disease diagnoses. The codes have the co-occurrence relation without order. For example, in a patient record, the chronic kidney disease is recorded after a cold record, but we can't conclude that the patient didn't have chronic kidney disease before he caught a cold. Two diagnoses have an uncertain time

*Correspondence: sun_chenxi@pku.edu.cn
[1] School of Electronics Engineering and Computer Science, Peking University, No. 5 Yiheyuan Road, Beijing 100871, People's Republic of China
Full list of author information is available at the end of the article

relation. We call such issues as the co-occurrence relation, such as complication, causation, and continuity. Thus, EHR has both the time relation and the co-occurrence relation.

Medical tasks such as disease prediction [1–3], concept representation [4, 5], and patient typing [6–8] are essential for personalized healthcare and medical research. Nevertheless, the tasks are challenging for physicians, considering the complex patient states, the amount of diagnosis, and the real-time requirement. Thus, a data-driven approach by learning from large accessible EHRs is the desiderata.

In recent years, the Deep Learning (DL) model has made remarkable achievements due to its strong learning ability and flexible architecture [9–13]: some DL methods can model the sequential time relation of medical data. For example, RETAIN [3] utilizes gated recurrent unit (GRU) [14, 15] to predict medical events, Dipole [1] uses Bidirectional RNN (BRNN) [16] to integrate the information in the past and the feature, and T-LSTM [8, 17] injects the time decay effect to handle irregular time intervals. Using these methods, the EHR structure is modeled as Fig. 1b; Some DL methods can model the co-occurrence relation of medical data. For example, Word-2Vec [18, 19], Med2Vec [4], and MiME [5] model the medical relations to better express the original data by the idea of representation learning [20–23]. Using these methods, the EHR structure is modeled as Fig. 1c.

However, no method can model both relations simultaneously. Because t there is a conflict between the two relations: The time relation makes data distributed longitudinally but the co-occurrence relation makes data distributed bipartite graph-like. If considering both these two relations, the EHR structure is shown in Fig. 1d.

Meanwhile, in the real-world application, the data-driven method is required to be interpretable to facilitate the use of doctors [24–26]. However, the DL method is the black-box model which is troubled by poor interpretability [27–32].

To address the above issues, in this work, we define EHR as the hierarchical co-occurrence sequence and propose a novel model called Time-aware and Co-occurrence-aware Network (TCoN). TCoN can not only model the two relations simultaneously but also has the ability of interpretation. TCoN has the pre-train and fine-tune mechanism for the imbalanced data and is more accurate than all baselines in medical prediction tasks.

## Materials and methods

In this section, we first introduce the MIMIC-III dataset and the data preprocessing process. Then, we describe the proposed methods in detail.

### Dataset description and preprocessing

MIMIC-III is a freely accessible de-identified medical dataset, developed and maintained by the Massachusetts Institute of Technology Laboratory for Computational Physiology [33]. Based on MIMIC-III dataset, we selectively extract data and form three data sets:

#### *Overall dataset*

We extract records with more than one visit from MIMIC-III. The new dataset comprises 19,993 hospital admissions of 7537 patients and 260,326 diagnoses with 4,893 unique codes defined by the International Classification of Diseases-9 version (ICD-9). For one patient, the visit number is 2.66 on average. For one visit, the code number is 13.02 on average and up to 39.

#### *Sepsis dataset*

Following the latest sepsis 3.0 definition [34], we extract 1232 sepsis patients whose SOFA is greater than or equal to 2.

#### *Heart failure dataset*

According to ICD-9 code, we extract 1608 heart failure patients who have diagnoses of 428.x code.

In sepsis dataset and heart failure dataset, the extracted data is the records for the first time that these two diagnoses appear. And these two datasets are imbalanced. The detailed statistic is shown in Table 1.

### Problem formulation

**Definition 1** (*Electronic Health Record* | *EHR*) EHR is the hierarchical co-occurrence time sequence data. It consists of a set of records $R = \{r_i | i = 1, \ldots, M\}$ with $M$ patients $P$. Each record $r_i$ has a visit sequence $V = \{v_i | i = 1, \ldots, N\}$ mapped in time. For each

(See figure on next page.)

**Fig. 1** The data structure of EHR based on different methods. **a** Original EHR data structure. **b** EHR data structure based on time relation. **c** EHR data structure based on co-occurrence relation. **d** Data relation under our TCoN model. The data form **b** arranges codes in a random order, but different sequences have different effects on results. For example, the sequence 'heart disease —> influenza —> coronary' has closer relation between 'heart disease' and 'influenza' than the sequence 'heart disease —> coronary —> influenza'. The data form **c** can make every two codes have the equal relation, but if 'heart disease', 'atrial fibrillation' and 'diabetes' are in three different visits, the equal relation will fail as there are different time intervals among them. The data form **d** is the combination. it describes both the equal code relation in the same visit and the time relation in different visits

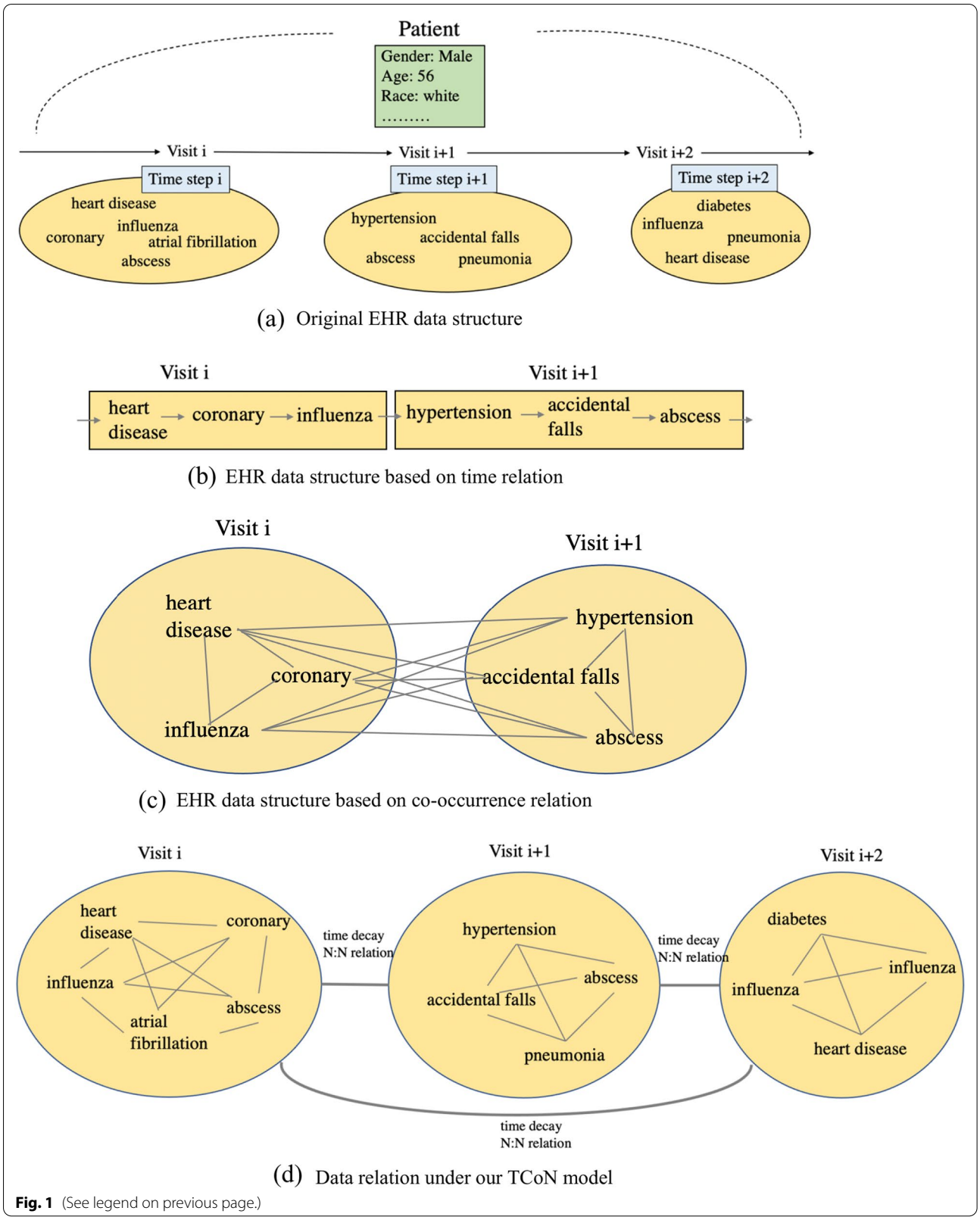



**Fig. 1** (See legend on previous page.)

**Table 1** Statics of extracted MIMIC-III dataset

| Sepsis codes | SOFA ≥ 2 |
|---|---|
| Heart failure codes | 428.x |
| Avg. rate of in-hospital mortality | 12.58% (5854/46,520) |
| Avg. rate of readmission to ICU | 16.20% (7537/46,520) |
| Avg. rate of sepsis | 6.16% (1232/19,993) |
| Avg. rate of heart failure | 8.04% (1608/19,993) |

$v_i$, it contains a time stamp $t_i$ and many codes $c_i = \{c_{ij}|c_{ij}\varepsilon C, j = 1, \ldots, J\}$. $C$ is a diagnoses database. Meanwhile, the demographic information $I$ is recorded to patients $P$.

**Definition 2** (*Medical prediction tasks*) They use a set of medical records $R$ to predict the specific target $Y = \{y_1, y_2, \ldots y_n\}$. If $n = 2$, it is a two-classification task. If $n > 2$, it is a multi-classification task. The prediction task is $f_p : R \rightarrow Y$.

**Definition 3** (*Interpretation Path*) Interpretation uses the correlations $\mathcal{R}$ of medical pairs $Q$ to build an Interpretation Path $\mathcal{P}$. $Q$ is a set of tuple $\{(a, b)|a\varepsilon C \cup V, b\varepsilon C \cup V \cup P \cup Y\}$, the pair correlation is $a\mathcal{R}b$, and $\mathcal{P} = c_1 \rightarrow^{\mathcal{R}_1} c_2 \rightarrow^{\mathcal{R}_2} \ldots \rightarrow^{\mathcal{R}_{n-1}} c_n \rightarrow^{\mathcal{R}_n} prediction$ interprets how TCoN predicts.

### Analysis strategy

**Task 1** (Mortality prediction). To predict if the patient will die during the hospitalization.

**Task 2** (Readmission prediction). To predict if the patient will be hospitalized again.

**Task 3** (Disease prediction). Two disease prediction tasks: Sepsis and heart failure. Early diagnose is critical for improving patients' outcome [35].

**Task 4** (Next diagnoses prediction). To predict the diagnoses of the patient in the next admission.

Note that Task 1, 2, 3 are binary classification tasks and Task 3 is a multi-classification task.

**Evaluation 1** (AUC-ROC). The area under the curve of the True Positive Rate (TPR) and the False Positive Rate (FPR). TN, TP, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

$$\begin{aligned} TPR &= \frac{TP}{TP + FN} \\ FPR &= \frac{FP}{TN + FP} \end{aligned} \tag{1}$$

**Evaluation 2** (PR-AUC). The area under the curve of Precision (P) and Recall (R). It is a better measure for imbalanced data [36].

$$\begin{aligned} P &= \frac{TP}{TP + FP} \\ R &= \frac{TP}{TP + FN} \end{aligned} \tag{2}$$

**Evaluation 3** (Accuracy@k). The probability of the positive predictions in top-k prediction values. It is the evaluation metric of multi-classification tasks.

$$Accuracy@k = \frac{\#\ of\ \ true\ positive\ in\ top\ k}{\min(k, |c_t|)} \tag{3}$$

### TCoN model structure

As shown in Fig. 2, our TCoN model contains the code block and the visit block: The code block is implemented by Co-occurrence-aware Self-attention (CS-attention); The visit block is implemented by Time-aware Gated Recurrent Unit (T-GRU); Two blocks are connected by Attention connection.

#### *CS-attention*

Self-attention [32] in natural language processing considers the semantic and grammatical relations between different words in sentences. For each input, it has three vectors, Query (Q), Key (K), and Value (V). The multi-head self-attention is designed as:

$$\begin{aligned} &Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{D_k}}\right)V \\ &MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \\ &head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{4}$$

In this work, we redesign the self-attention as CS-attention (Eq. 5) to deal with the relations of EHR codes. CS-attention has two different heads—Local Head and Global Head. The local head learns the co-occurrence relations between every two codes in the same visit. A code is affected by the other codes equally. The global head learns the co-occurrence relations between every two codes in different visits. A code has different effects from the other codes according to different time intervals between visits. These two types of heads can learn a new representation $\tilde{C}$ of each code $C$ by its neighbors $C_{nb}$.

$$\tilde{C} = Attention(C) = softmax\left(\frac{Q_1K_1^T}{\sqrt{d_k}}, \frac{(Q_2K_2^T)^T T}{\sqrt{d_k}}\right)(C, C_{nb}) \tag{5}$$

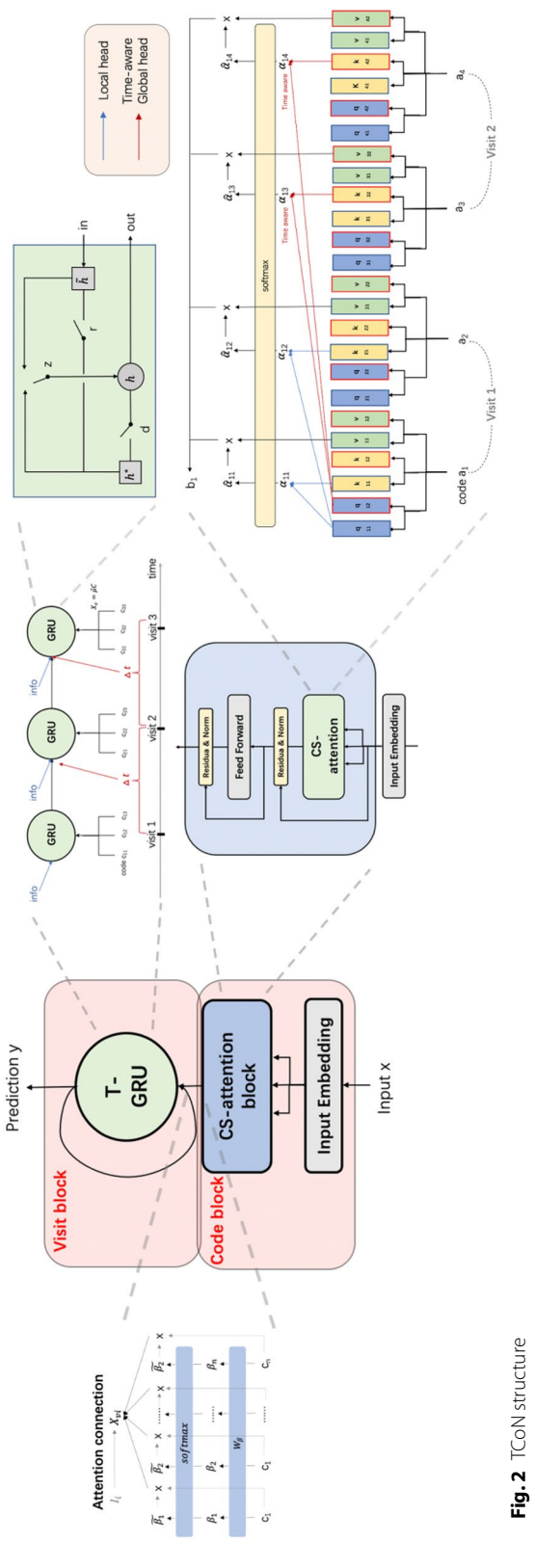



**Fig. 2** TCoN structure

$C$ is the original matrix of input codes. $\tilde{C}$ is the new representation matrix. $Q_i, K_i, V_i$ is same as Eq. (4). $i = 1$ represents the local head and $i = 2$ represents the global head. $d_k$ is the dimension for $Q$ and $K$. $T$ is the time decay function $g(\Delta t)$ in Eq. 4. Both the number of local head and global head can be change.

### *T-GRU*

As shown in Eq. 3, T-GRU comprises an update gate $z_t$ and a reset gate $r_t$. They control the extent to which the previous state $h_{t-1}$ is brought into the current state $h_t$ and how far the previous state is brought into the current candidate state $\tilde{h}_t$. For modeling the time irregularity, we build a time gate $d_t$. This gate takes time interval into account and control delivered information from the previous visit to the current visit by time decay function $g(\Delta t)$. The time decay function can determine how much the history state can be injected into the current unit. In Eq. 3, $x_t$ is the current input data, $W, U, b$ are parameters. The output is the current state $h$.

$$\begin{aligned}
&\text{Time gate}: \quad d_t = g(\Delta t)\\
&\text{Time decayed history state} \quad h^*_{t-1} : h^*_{t-1} = h_{t-1} \cdot d_t\\
&\text{Update gate } z_t: \quad z_t = \sigma\left(x_t \cdot W_z + h^*_{t-1} \cdot U_z + b_z\right)\\
&\text{Reset gate } r_t: \quad r_t = \sigma\left(x_t \cdot W_r + h^*_{t-1} \cdot U_r + b_r\right)\\
&\text{Candidate state } \tilde{h}_t: \quad \tilde{h}_t = \left(x_t \cdot W_h + h^*_{t-1} \cdot r_t \cdot U_z + b_h\right)\\
&\text{Current state } h_t: \quad h_t = z_t \cdot h^{*T}_{t-1} + (1 - z_t) \cdot \tilde{h}_t
\end{aligned} \tag{6}$$

We propose three time decay functions (Eq. 7). $\Delta t$ is the time interval between two visits, $\alpha$ is the decay rate. When $\alpha = 1$, the exponential form is more suitable for the small elapsed time, the logarithmic form is more suitable for the large elapsed time, and the reciprocal form is a compromise.

$$\begin{aligned}
&\text{Reciprocal form} \quad g(\Delta t) = \frac{1}{1 + \alpha \Delta t}\\
&\text{Logarithmic form} \quad g(\Delta t) = \frac{1}{\log(e + \alpha \Delta t)}\\
&\text{Exponential form} \quad g(\Delta t) = e^{-\alpha \Delta t}
\end{aligned} \tag{7}$$

### *Attention connection*

Between code block and visit block, we design the connection method (Eq. 8). Where $X_{vi}$ is the $i$th input of visit $v$, $C_i$ is the output matrix with each row for one $i$-th visit's code, $W_\beta$ is a parameter vector. When we

consider the demographic information $I$. The input will be a concatenation form: $X_{vi} = concate\left(\tilde{\beta}^T C_i, I_i\right)$.

$$\begin{aligned} \beta &= C_i W_\beta + b_\beta \\ \tilde{\beta} &= softmax(\beta) \\ X_{vi} &= \tilde{\beta}^T C_i \end{aligned} \tag{8}$$

Besides, we propose a method to interpret TCoN. It is achieved by the correlation values among codes, visits, and predictions.

#### *Interpretation path*

It is based on the correlations $\mathcal{R}$, containing two correlations: The code-code correlation is obtained from $\widehat{\alpha}$ of CS-attention. $\widehat{\alpha}_{ij}$ means the effect of code $j$ on code $i$, and large $\widehat{\alpha}_{ij}$ means that code $j$ could be the cause, complication, or early symptoms of code $i$; The code-visit correlation is obtained from $\beta$ of the Attention connection. Larger $\beta$ means the closer relation.

The interpretation path is a code sequence obtained by the reverse lookup starting with the prediction results. For a prediction $P$, the last visit is $v_n$. In $v_n$, we find the code $c_{ni}$ that contributed the most to $v_n$ according to $\beta$. For $c_{ni}$, we find the closest code $c_{(n-1)i}$ in visit $v_{n-1}$ according to the largest $\widehat{\alpha}_{*C_{ni}}$. Similarly, we find $c_{(n-2)i}, c_{(n-3)i}, \ldots c_{1i}$. So far, we find a path $c_{1i} \rightarrow \cdots \rightarrow c_{ni} \rightarrow P$. This path can be described: a disease $c_{1i}$ most likely infers $c_{2i}$, then $c_{2i}$ most likely infers $c_{3i}$, ... and $c_{(n-1)i}$ most likely infers $c_{ni}$, finally, $c_{ni}$ most likely causes $P$.

Finally, we apply a training method that enables TCoN to handle imbalanced data [37, 38].

#### *Pre-train and Fine-tune*

In the pre-train process, we apply an auto-encoder network $f_{ae}$ with a minimum loss (Eq. 9) for the unsupervised representation learning task. In the fine-tune process, we use parameters of the encoder layer as the initial parameters of TCoN when training by the prediction objective in Eq. (10). For TCoN, the input layer is represented by Eq. (8), Skip-connection is Eq. (12), layer normalization [29] is Eq. (13), and feed forward layer is Eq. (14).

$$L_{emb} = -\frac{1}{n}\sum_{i}^{n} x_i log f_{ae}(x_i) \tag{9}$$

$$L_{pre} = -\frac{1}{n}\sum_{i}^{n} y_i log f_{pre}(x_i) \tag{10}$$

$$a = emb(x) = ReLU(x \cdot W + b) \tag{11}$$

$$x' = RC(x) = x + f(x) \tag{12}$$

$$x' = \gamma \frac{x - \mu}{\sqrt{\sigma + \varepsilon}} + \beta \tag{13}$$

$$c = FF(b) = Relu(b \cdot W_1 + b_1) \cdot W_2 + b_2 \tag{14}$$

#### Complexity analysis

The self-attention-based algorithm is parallel, but the RNN-based algorithm is serial [32]. TCoN has both structures and they are connected in series. Thus, the complexity of TCoN is $O\left(n^2 \cdot d\right) = O\left(n^2 \cdot d\right) + O\left(n \cdot d^2\right)$. $d$ is the representation dimension and $n$ is the sequence length. $O\left(n^2 \cdot d\right)$ is the complex of CS-attention with $n^2$ for operations of every two inputs. $O\left(n \cdot d^2\right)$ is the complex of T-GRU with $d^2$ for sequential operation. In our data, the dimensionality $d$ is smaller than the data length $n$, so that the complex of TCoN is $O\left(n \cdot d^2\right)$.

## Results

### Experimental setup

For data, we right align the time series and use padding and masking to make them equal in length. Each code is represented by a one-hot vector with 4,893 dimensions (number of ICD-9 codes). Training, validation, and testing set is in 0.75:0.1:0.15 ratio.

For model, we set 2 local heads and 2 global heads. We choose $\alpha = 1$ logarithmic time decay with year as the decay unit. We apply Adam Optimizer [39] with $\alpha = 0.001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use the learning rate decay method $\alpha_{current} = \alpha_{initial} \cdot \gamma^{\frac{global\ step}{decay\ steps}}$ with decay rate $\gamma = 0.98$ and decay step = 2000 [40]. Before the prediction task, we carry out the pre-train step and use the early stop with 5 epochs. We use the fivefold cross-validation. The code implementation is publicly available at https://github.com/SCXsunchenxi/MTGRU

### Baselines

- Time-aware methods (RNN-based methods)
  - GRU [14]. It uses GRU to embed visits and make the final prediction.
  - T-LSTM [8]. It uses elapsed time weight to change previous memory in LSTM.
- Co-occurrence-aware methods (Word2Vec-based methods)

- Med2Vec [4]. It applies the skip-gram model and multi-layer perceptron to get the representation of codes and visits.
- Dipole [1]. It uses BRNN along with three attention mechanisms to measure the relation of different visits for the final prediction.

### Prediction results

TCoN predicts more accurately than all baselines. The results of binary classification (mortality, readmission, sepsis, and heart failure) and multi-classification (next diagnoses) are shown in Table 2(a, b). Baselines may not match EHR characteristics and partially model data features. For example, T-LSTM has the worst performance as it is not suitable for short visit sequences like MIMIC-III.

TCoN performs well on imbalanced datasets. In binary classification tasks, all datasets are imbalanced, especially the sepsis dataset (6.16%). But the results show that the more imbalanced the data, the greater the advantage of TCoN over baselines.

TCoN can accurately predict multiple diagnoses in the next admission. In the multi-classification task, we evaluate methods with k = 5, 15, 25, 35. As shown in Table 2b, as k increases, the accuracies of all methods decrease, but the advantage of our approach is still obvious.

### Model parameters experiments

We change the dimension of representation vector in hidden layers. The results in Fig. 3a show that TCoN performs better than other methods under all dimensions. Then, we set different numbers of heads for TCoN. Figure 3b shows that the number of heads = 2 is the key turning point.

### Case study of interpretation path

We choose a patient numbered 32,790 in MIMIC-III (a white man with 3 admission records and died at 80) to describe how TCoN produces the interpretation path. Figure 4a is the heat map of $\widehat{\alpha}$ for the death prediction. The diagnosis 'hypoxemia' contributes the most to the last admission as its weighted vector's norm is the biggest. For 'hypoxemia', the closest diagnosis is 'pulmonary collapse' with the biggest $\widehat{\alpha}_{*i} = 0.892$. For 'pulmonary collapse', the closest diagnosis is 'unspecified pleural effusion' with the biggest $\widehat{\alpha}_{*i} = 0.803$. And for 'unspecified pleural effusion', the closest diagnosis is 'unspecified sleep apnea' with the biggest $\widehat{\alpha}_{*i} = 0.782$. So far, an interpretation path 'unspecified sleep apnea —> unspecified pleural effusion —> pulmonary collapse —> Hypoxemia —> death' is found as shown in Fig. 4b.

Figure 4c shows cases of interpretation paths of sepsis prediction and heart failure prediction. Each path is the summary results by using the most frequent diagnosis. Thus, we find sepsis-related pre-diagnoses/symptoms, such as 'Fever', 'Chills', 'Immunity disorders', 'Anemia' and 'Coma'. And we find heart failure-related pre-diagnoses/symptoms, such as 'Ventricular fibrillation', 'Myocarditis', 'Coronary atherosclerosis' and 'Hypertension'.

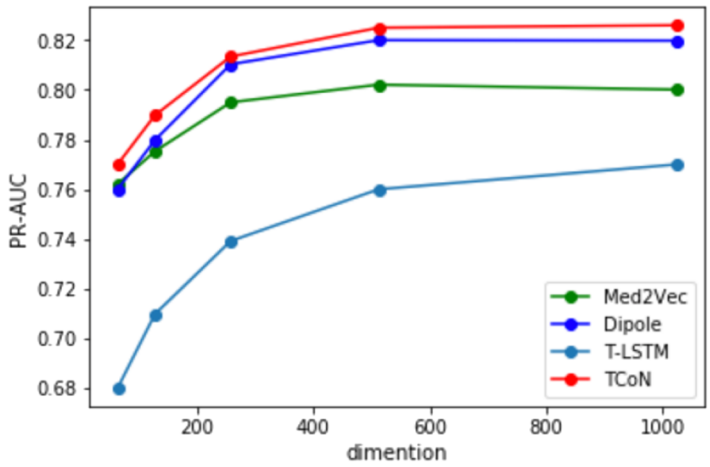
## Discussion

In recent years, deep learning (DL) technology has shown its superior performance in medical applications [41–44], such as medical image recognition [45] and

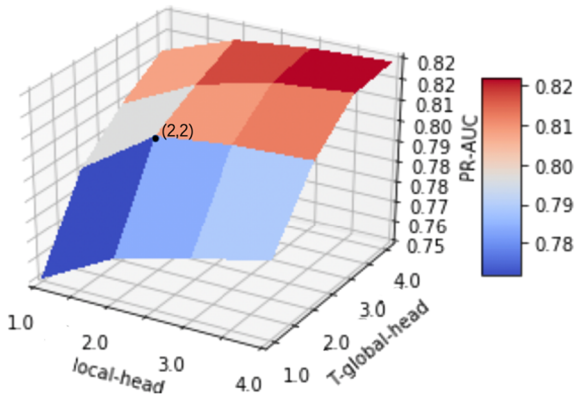
**Table 2** Prediction results of mortality, readmission, sepsis, heart failure and next diagnoses

| Method | Mortality | | Readmission | | Sepsis | | Heart Failure | |
|---|---|---|---|---|---|---|---|---|
| | ROC-AUC | PR-AUC | ROC-AUC | PR-AUC | ROC-AUC | PR-AUC | ROC-AUC | PR-AUC |
| (a) *Results of binary classification prediction* | | | | | | | | |
| GRU | 0.7902 | 0.7400 | 0.7023 | 0.6713 | 0.6202 | 0.6063 | 0.6525 | 0.6187 |
| Med2Vec | 0.8025 | 0.7950 | 0.7125 | 0.6833 | 0.8211 | 0.7943 | 0.7225 | 0.7101 |
| Dipole | 0.8133 | 0.8103 | 0.7341 | 0.7243 | 0.8001 | 0.7823 | 0.7067 | 0.6923 |
| TLSTM | 0.7893 | 0.7392 | 0.7256 | 0.7023 | 0.6432 | 0.6189 | 0.7432 | 0.6033 |
| TCoN | 0.8224 | 0.8134 | 0.7403 | 0.7278 | 0.8433 | 0.8233 | 0.7698 | 0.7313 |

| | Accuracy@5 | Accuracy@15 | Accuracy@25 | Accuracy@35 |
|---|---|---|---|---|
| (b) *Multi-classification result of next diagnoses prediction* | | | | |
| GRU | 0.7723 | 0.6298 | 0.5801 | 0.4523 |
| Med2Vec | 0.8025 | 0.7061 | 0.6250 | 0.5025 |
| Dipole | 0.8043 | 0.6514 | 0.6012 | 0.5044 |
| TLSTM | 0.7833 | 0.6367 | 0.5814 | 0.4515 |
| TCoN | 0.8398 | 0.7223 | 0.6577 | 0.5113 |




(a) The classification accuracy under different representation dimensions. Specifically, parameters are: $d$ and $m$ of TCoN; $m$ and $n$ of Med2Vec, $m$ of Dipole, and $m$ of T-LSTM.



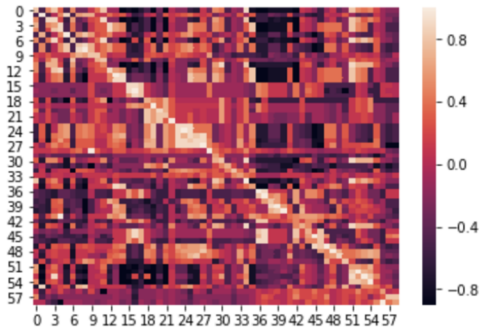

(b) The classification accuracy under different number of heads of TCoN

**Fig. 3** The classification accuracy under different model parameters

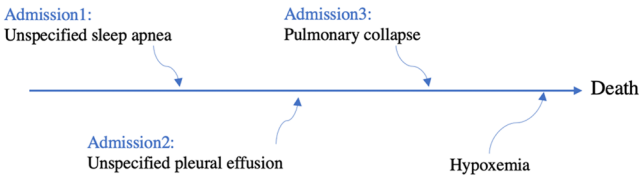
medication recommendations [46]. And many methods have achieved good performance for specific disease prediction, such as Alzheimer's disease [47], sepsis [48], and heart disease [49, 50]. However, most of them pursue the task accuracy but ignoring the interpretability. DL-based approaches are black-box models, which is not easy to understand for non-professionals, especially doctors without artificial intelligence backgrounds. Thus, the explainable DL method is needed. This study aims at this problem and puts forward a solution, interpretation path, to make the predictions explainable.

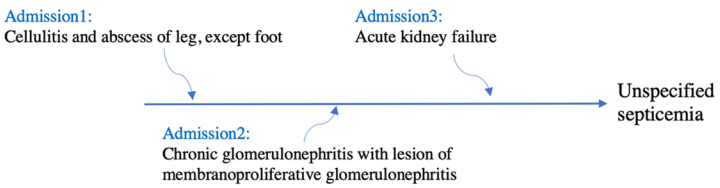
In EHR, the patient's records are irregular in time due to the unpredictability of the diseases and inevitable data loss. The current disease could be more closely related to the disease a week ago than the disease a year ago [8, 9]. Thus, the time perception mechanism is needed. This study aims at this issue and proposes a time gate to explicitly learn the irregular time information by the time decay function.

The experiments show that using two kinds of head for relations of inter-visit and intra-visit is necessary. The difference between these two relations is not just the time interval, but also the pathology. We emphasize the code relations are more likely to be complications in the same visit, but causations and continuities among different visits. For example, in our experiments, the relation of 'diabetes' with 'cellulitis and abscess of legs' in one visit is more prone to be a short-term complication, but the




(a) The heat map of $\hat{\alpha}$ in death prediction of patient 32790.




(b) The interpretation path of death prediction and 'unspecified septicemia' of patient 32790.

| Admission No. | Most Relevant Diagnosis | Statistics |
|---|---|---|
| N-4 | Fever, Chills (ICD-9: 780.6) | 6/59 |
| N-3 | Immunity disorders (ICD-9: 279) | 59/65 |
| N-2 | Anemia (ICD-9: 283.9, 285.1, 283.19, 773.2) | 65/323 |
| N-1 | Coma (ICD-9:780.01, 572.2) | 323/1,232 |
| N | Sepsis | 1,232 |

| Admission No. | Most Relevant Diagnosis | Statistics |
|---|---|---|
| N-4 | Hypertension (ICD-9:997.91, 401. 0-401.9) | 8/21 |
| N-3 | Coronary atherosclerosis (ICD-9:414.0) | 21/102 |
| N-2 | Myocarditis (ICD-9:422,398,429) | 102/621 |
| N-1 | Ventricular fibrillation (ICD-9: 427.4) | 621/1,606 |
| N | Heart failure | 1,606 |

(c) The interpretation paths for sepsis prediction and heart failure prediction.

**Fig. 4** Interpretation path

relation of 'diabetes' and 'long-term use of insulin' in two different visits is more prone to be causation. Thus, for each patient, we can give a disease association graph. The weight of the edges between two diagnoses in the same admission represents the adjoint coefficient, and the weight of the edges between two diagnoses in different admissions represents the causal coefficient. Figure 5 shows the diagnosis graph case of patient 32,790.

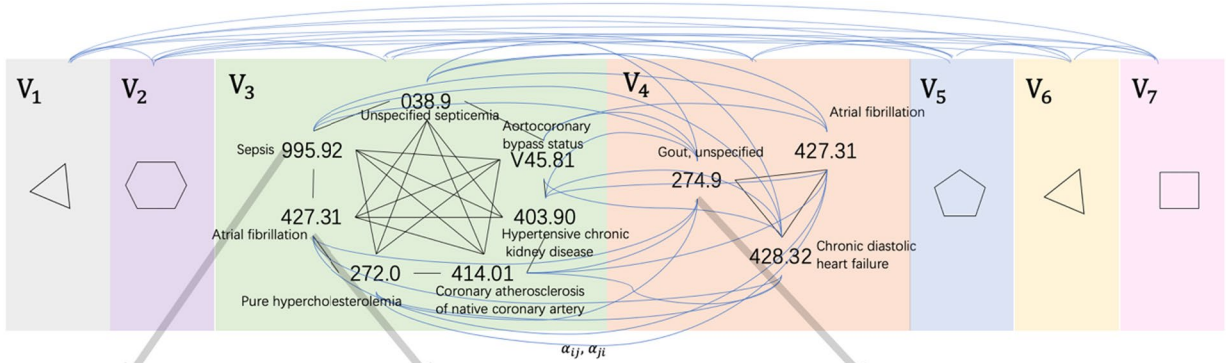
The interpretation path is not symmetrical, which means $\widehat{\alpha}_{ij} \neq \widehat{\alpha}_{ji}$. $\widehat{\alpha}_{ij}=\frac{\#\ of\ i-j\ occurrences}{\#of\ i\ occurrences}$ and $\widehat{\alpha}_{ji} = \frac{\#\ of\ i-j\ occurrences}{\#\ of\ j\ occurrences}$, they have different denominators. For example, code $i$, $j$, $k$ represent the diagnoses of 'malaria', 'fever', 'periodic cold fever' respectively. In our experiment, $i$ is mostly accompanied by $j$ as $\widehat{\alpha}_{ij} = 0.762$. But $j$ is not always accompanied by $i$ as $\widehat{\alpha}_{ji} = 0.023$. It is mostly accompanied by code $k$ with $\widehat{\alpha}_{ki} = 0.701$. Comparing $\widehat{\alpha}_{ji}$ and $\widehat{\alpha}_{ki}$, the results show that 'periodic cold fever' is a better explanation for 'malaria' than 'fever'. In research [51], 'periodic cold fever' is a special clinical manifestation of 'malaria' and there are very few other diseases with this symptom. It illustrates that our interpretable method can explain the results by reflecting the relation (such as complication, causation, and continuity) between the diagnoses and $\widehat{\alpha}_{*i}$ is a more important standard to find the maximum co-occurrence code for $i$ than $\widehat{\alpha}_{i*}$.

In medical applications, the data is usually imbalanced. The normal state of patients is the majority, while the disease records may be the small sample. But the small sample is more important for the disease prediction. Thus, a DL model should be robust on the imbalanced dataset. In this paper, our pre-train and fine-tune framework can help.

Further, there is room for further improvement. The current modeling method is based on pure EHRs data. Integrating prior information will make the results of the data relation modeling and medical prediction more accurate and reasonable. The available method is knowledge graph embedding based on ICD code. Besides, more data in EHRs such as doctor notes, medications, and laboratory tests can be used for better performance. Future work will focus on these aspects.

## Conclusion

The data-driven medical prediction method based on interpretable deep learning is essential for healthcare management. In this paper, we propose an




| code | $\alpha_{ij}$ rank / $\alpha_{ji}$ rank | code |
|---|---|---|
| 995.92 | 3 / 7 | 427.31 |
| 995.92 | 2 / 5 | 272.0 |
| 995.92 | 5 / 5 | 414.01 |
| 995.92 | 6 / 4 | 403.90 |
| 995.92 | 1 / 1 | 038.9 |
| 995.92 | 4 / 6 | V45.81 |
| 995.92 | 9 / 8 | 274.9 |
| 995.92 | 8 / 7 | 428.32 |
| 995.92 | 7 / 6 | 427.31 |

| code | $\alpha_{ij}$ rank / $\alpha_{ji}$ rank | code |
|---|---|---|
| 427.31 | 6 / 3 | 995.92 |
| 427.31 | 1 / 3 | 272.0 |
| 427.31 | 5 / 2 | 414.01 |
| 427.31 | 8 / 7 | 403.90 |
| 427.31 | 7 / 5 | 038.9 |
| 427.31 | 2 / 1 | V45.81 |
| 427.31 | 9 / 3 | 274.9 |
| 427.31 | 4 / 7 | 428.32 |
| 427.31 | 3 / 5 | 427.31 |

| code | $\alpha_{ij}$ rank / $\alpha_{ji}$ rank | code |
|---|---|---|
| 274.9 | 4 / 5 | 427.31 |
| 274.9 | 2 / 5 | 272.0 |
| 274.9 | 5 / 5 | 414.01 |
| 274.9 | 6 / 7 | 403.90 |
| 274.9 | 8 / 6 | 038.9 |
| 274.9 | 7 / 7 | V45.81 |
| 274.9 | 9 / 5 | 995.92 |
| 274.9 | 3 / 4 | 428.32 |
| 274.9 | 1 / 3 | 427.31 |

**Fig. 5** Diagnosis graph of patient 32,790. Patient 32,790 has 7 visits, shown as 7 blocks with different colors. For clarity, we only show the ICD-9 code of diagnoses in visit 3, 4. The black line represents the code relations in the same visit, the blue lines represent the code relations in different visits. The relation closeness is measured by the edge weight. Every edge between code $i$ and $j$ has two weights $\alpha_{ij}$ and $\alpha_{ji}$. Three tables record the relations between code 995.52 (sepsis), code 427.31 (atrial fibrillation), code 274.9 (gout) with other codes

interpretable Time-aware and Co-occurrence-aware Network (TCoN) for data modeling and medical prediction. It can perceive hierarchical data structures with the time relation and the co-occurrence relation, give an interpretation path to explain the prediction, and build a diagnosis graph for every patient. The experiments show that TCoN outperforms the state-of-the-art methods.

#### Abbreviations

EHR: Electronic health record; ICD: International classification of diseases; WHO: World Health Organization; DL: Deep learning; RL: Representation learning; RNN: Recurrent neural network; CNN: Convolutional neural network; BRNN: Bidirectional recurrent neural network; LSTM: Long short-term memory; GRU: Gated recurrent unit; TCoN: Time-aware and co-occurrence-aware deep learning network; CS-attention: Co-occurrence-aware self-attention; T-GRU : Time-aware gated recurrent unit; AUC-ROC: The area under the curve of receiver operating characteristic; PR-AUC: The area under curve of precision-recall; TP: True positives; TN: True negatives; FP: False positives; FN: False negatives; TPR: True positive rate; FPR: False positive rate.

#### Acknowledgements

This paper is dedicated to those who want to fight COVID-19.

#### Authors' contributions

CS and HL conceptualized the idea. HL initialized and supervised the project. CS collected data, implemented the experiments, and drafted the manuscript. HD reviewed the manuscript and implemented the additional experiments. All authors provided a critical review of the manuscript and approved the final draft for publication.

#### Funding

This work was supported by the National Natural Science Foundation of China (No. 62172018, No. 62102008) and the National Key Research and Development Program of China under Grant 2021YFE0205300 to collect and process data and publish the paper.

#### Availability of data and materials

The code implementation is publicly available at https://github.com/SCXsunchenxi/MTGRU. The data is at https://mimic.physionet.org.

## Declarations

#### Ethics approval and consent to participate

MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). https://doi.org/10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635

#### Consent for publication

Not applicable.

#### Competing interests

No financial competing interests.

#### Author details

[1]School of Electronics Engineering and Computer Science, Peking University, No. 5 Yiheyuan Road, Beijing 100871, People's Republic of China. [2]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, People's Republic of China. [3]The Aviation Industry Corporation of China, Ltd, Chengdu Aircraft Design & Research Institute, Chengdu 610041, People's Republic of China.

Received: 8 October 2020 Accepted: 18 October 2021
Published online: 02 November 2021

#### References

1. Ma F, Chitta R, Zhou J, You Q, Sun T, Gao J. Dipole: diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In: SIGKDD; 2017.
2. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. In: ICLR; 2016.
3. Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In: NIPS; 2016. p. 3504–3512.
4. Choi E, Bahadori MT, Searles E, Coffey C, Thompson M, Bost J, Tejedor-Sojo J, Sun J. Multi-layer representation learning for medical concepts. In: SIGKDD; 2016. p. 1495–1504.
5. Choi E, Xiao C, Stewart WF, Sun J. Mime: multilevel medical embedding of electronic health records for predictive healthcare. In: NIPS; 2018.
6. Li H, Li X, Jia X, Ramanathan M, Zhang A. Bone disease prediction and phenotype discovery using feature representation over electronic health records. In: ACM-BCB; 2015.
7. Che Z, Kale D, Li W, Bahadori MJ, Liu Y. Deep computational phenotyping. In: SIGKDD; 2015. p. 507–516.
8. Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J. Patient subtyping via time-aware lstm networks. In: SIGKDD; 2017.
9. Pham T, Tran T, Phung D, Vankatesh S. DeepCare: a deep dynamic memory model for predictive medicine. arxiv: 1602.00357v1.
10. Razavian N, Sontag D. Temporal convolutional neural networks for diagnosis from lab tests. CoRRabs/1511.07938, 2015.
11. Cheng Y, Wang F, Zhang P, Hu J. Risk prediction with electronic health records: a deep learning approach. In: SDM; 2016. p. 432–440.
12. Che Z, Cheng Y, Sun Z, Liu Y. Exploiting convolutional neural network for risk prediction with medical feature embedding. CoRR abs/1701.0747, 2017.
13. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25(2):1097–105.
14. Chung J, Gulcehre C, Cho KH, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. Eprint Arxiv, 2014.
15. Schmidhuber J. Learning complex, extended sequences using the principle of history compression. Neural Comput. 2014;4(2):234–42.
16. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Signal Process. 2002;45(11):2673–81.
17. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
18. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: NIPS; 2013.
19. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: ICML; 2013.
20. Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilistic language model. In: JMLR; 2003.
21. Mnih A, Hinton GE. A scalable hierarchical distributed language model. In: NIPS; 2009.
22. Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning. In: ACL; 2010.
23. Wang Y, Yuan Y, Ma Y, et al. Time-dependent graphs: definitions, applications, and algorithms. Data Sci Eng. 2019;4:352–66.
24. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: ICLR; 2015.
25. Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. In: EMNLP; 2015. p. 1412–1421.
26. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel RS, Bengio Y. Show, attend and tell: neural image caption generation with visual attention. In: ICML, 2015.
27. You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. In: CVPR; 2016. p. 4651–4659.
28. Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention. In: ICLR; 2015.
29. Ba JL, Kiros JR, Hinton GE. Layer CoRR abs/1607.06450, 2016.
30. Chorowski JK, Bah-danau D, Serdyuk D, Cho K, Bengio Y. Attention-based models for speech recognition. In: NIPS; 2015. p. 577–585.
31. Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P. Teaching machines to read and comprehend. In: NIPS; 2015. p. 1693–1701.

32. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser K. Attention is all you need. In: NIPS, 2017.
33. Johnson A, Pollard T, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi L, Mark R. Mimic-III, a freely accessible critical care database. Sci Data. 2016;3(1):1–9.
34. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA. 2016;315:8.
35. Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, Lemeshow S, Osborn T, Terry KM, Levy MM. Time to treatment and mortality during mandated emergency care for sepsis. N Engl J Med. 2017;376(23):2235–44.
36. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):e0118432.
37. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional trans-formers for language understanding. arXiv preprint arXiv: 1810.04805, 2018.
38. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language under-standing by generative pre-training; 2018.
39. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: ICLR (Poster); 2015.
40. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: ICLR; 2019.
41. Pham T, Tran T, Phung D, Vankatesh S. DeepCare: a deep dynamic memory model for predictive medicine. arxiv: 1602.00357v1, 2016.
42. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling; 2018.
43. Adam G, Rampášek L, Safikhani Z, et al. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Onc. 2020;4:19.
44. Jalali A, Lonsdale H, Do N, et al. Deep learning for improved risk prediction in surgical outcomes. Sci Rep. 2020;10:9289.
45. Wang S, Wang S, Zhang S, Fan F, He G. Research on recognition of medical image detection based on neural network. IEEE Access. 2020;8:94947–55.
46. Shang J, Xiao C, Ma T, Li H, Sun J. GAMENet: graph augmented MEmory networks for recommending medication combination. In: AAAI; 2019. p. 1126–1133.
47. Dong Q, Zhang J, Li Q, Thompson PM, Caselli RJ, Ye J. Multi-task dictionary learning based on convolutional neural networks for longitudinal clinical score predictions in Alzheimer's disease. In: HBAI@IJCAI; 2019. p. 21–35.
48. Raghu A, Ko-morowski M, Singh S. Model-based reinforcement learning for sepsis treatment. In: ML4H workshop, NeurIPS; 2018.
49. Choi E, Schuetz A, Stewart WF, Sun J, Using recurrent neural network models for early detection of heart failure onset. J Am Med Inf Assoc. 2017.
50. Zhou Y, Hong S, Shang J, Wu M, Wang Q, Li H, Xie J. K-margin-based residual-convolution-recurrent neural network for atrial fibrillation detection. IJCAI. 2019; 6057–6063.
51. Peters D, Gray R, Jef VDE, et al. When is fever malaria? Lancet. 1992;339(8794):691.


## Publisher's Note