Ma *et al.*
*Health Information Science and Systems (2023) 11:7*
https://doi.org/10.1007/s13755-023-00210-5

Health Information Science and Systems

RESEARCH

# Knowledge and data-driven prediction of organ failure in critical care patients

Xinyu Ma[1], Meng Wang[1], Sihan Lin[2], Yuhao Zhang[3], Yanjian Zhang[1], Wen Ouyang[2] and Xing Liu[2*]

## Abstract

**Purpose:** The early detection of organ failure mitigates the risk of post-intensive care syndrome and long-term functional impairment. The aim of this study is to predict organ failure in real-time for critical care patients based on a data-driven and knowledge-driven machine learning method (DKM) and provide explanations for the prediction by incorporating a medical knowledge graph.

**Methods:** The cohort of this study was a subset of the 4,386 adult Intensive Care Unit (ICU) patients from the MIMIC-III dataset collected between 2001 and 2012, and the primary outcome was the Delta Sequential Organ Failure Assessment (SOFA) score. A real-time Delta SOFA score prediction model was developed with two key components: an improved deep learning temporal convolutional network (S-TCN) and a graph-embedding feature extraction method based on a medical knowledge graph. Entities and relations related to organ failure were extracted from the Unified Medical Language System to build the medical knowledge graph, and patient data were mapped onto the graph to extract the embeddings. We measured the performance of our DKM approach with cross-validation to avoid the formation of biased assessments.

**Results:** An area under the receiver operating characteristic curve (AUC) of 0.973, a precision of 0.923, a NPV of 0.989, and an F1 score of 0.927 were achieved using the DKM approach, which significantly outperformed the baseline methods. Additionally, the performance remained stable following external validation on the eICU dataset, which consists of 2,816 admissions (AUC = 0.981, precision = 0.860, NPV = 0.984). Visualization of feature importance for the Delta SOFA score and their relationships on the basic clinical medical (BCM) knowledge graph provided a model explanation.

**Conclusion:** The use of an improved TCN model and a medical knowledge graph led to substantial improvement in prediction accuracy, providing generalizability and an independent explanation for organ failure prediction in critical care patients. These findings show the potential of incorporating prior domain knowledge into machine learning models to inform care and service planning.

**Keywords:** Organ failure, Intensive care, Medical knowledge graph, Explainable model, Temporal convolutional network, Knowledge graph embedding

## Introduction

Multiple organ failure syndrome (MOFS) is a serious complication in intensive care unit (ICU) patients, with mortality rates ranging from 30 to 100% depending on the degree of organ failure [1]. The Sequential Organ Failure Assessment (SOFA) score was developed to describe the degree of organ failure over time and has been widely validated as the main criterion in the diagnosis of sepsis [2]. Identifying patients at risk of organ failure prior to any clinical manifestations would have a significant impact on the overall mortality and cost burden of sepsis and MOFS [3].

The availability of massive medical data and advances in deep learning methods have provided new opportunities for identifying high-risk patients to assist ICU clinical

*Correspondence: xingxingmail@csu.edu.cn
[2] Department of Anesthesiology, Third Xiangya Hospital, Central South University, Changsha 410013, People's Republic of China
Full list of author information is available at the end of the article

decision making. Models have developed based on ICU data to predict organ failure for patients; unfortunately, most existing deep learning methods [4–7] ignore the rich information embedded in medical knowledge, such as the correlations among different medical features and the importance of these features for different organ failure statuses, which leads to a bottleneck in model performance. To address these problems, we proposed an approach that incorporates prior medical knowledge and focuses on different features for different organ failure statuses to achieve high real-time predictive accuracy.

Most deep learning methods have not achieved significant deployment in clinical practice because most of them lack an explanation of how they operate, which is of paramount importance for clinical doctors to make their own informed and confident decisions [8]. Existing explanations [9, 10] provide insight into model training and parameters, i.e., concentrate on the specific features and the associated data which are responsible for a particular prediction or outcome. Independent explanations, which can be derived from external medical and biological evidence, theories, background knowledge, etc., are still lacking for deep learning models. A medical knowledge graph (MKG) is a reasonable approach for obtaining an independent explanation, carrying massive and structured medical background knowledge that is convenient for computer calculation. MKGs are applicable to a variety of downstream tasks, and the knowledge that can be extracted from them is easily understood by doctors [11, 12].

In this study, we propose a data- and knowledge-driven approach (called DKM) that takes full advantage of medical knowledge, clinical data, and deep learning algorithms to predict organ failure and provide explanations independent of the algorithms and clinical data.

## Methods

### Study design and data source

We trained and validated a prediction model based on the retrospective analysis of two nonoverlapping ICU databases:

(1) The MIMIC-III, an integrated, deidentified, comprehensive clinical dataset containing information on all patients admitted to the ICUs of Beth Israel Deaconess Medical Center in Boston, MA, from June 1, 2001, to October 31, 2012. There were 49,785 distinct hospital admissions to the ICUs during the study period.
(2) The eICU, a nonprofit program that was established by Philips and governed by customers, is a platform built from a repository of data used to advance knowledge of critical and acute care. It contains more than 3.3 million admissions from 2003 to 2016 in 459 ICUs across the United States.

The MIMIC-III dataset was used for model training and internal validation, and the eICU was used for external validation. Both sets of data included symptoms, vital signs, laboratory tests, International Classification of Disease (ICD-10) clinical diagnosis, and perioperative information. The study protocols conformed to the ethical guidelines of the 1975 Declaration of Helsinki. The reporting of studies conforms to the Transparent Reporting of a multifeature prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement [13] (Supplementary Table 1). The institutional review board (IRB) of 3rd Xiangya Hospital approved these studies (No. 21117).

#### *Patient cohorts*

From both the MIMIC-III and eICU datasets, we included adult patients aged $\geq$ 18 years. Patients were excluded if they did not have a SOFA score or vital sign data (i.e., heart rate, systolic blood pressure, diastolic blood pressure, mean arterial pressure, respiratory rate, or body temperature) within the first day of ICU admission or if they had error values (e.g., age > 120 years). The flowchart of patient selection is shown in Fig. 1.

### Data extraction and preprocessing

For both datasets, the ICU admissions of patients who met the above inclusion and exclusion criteria were included. The SOFA score was used to quantitatively describe the degree of organ failure in critical illness according to the function of six different organ systems (i.e., respiratory, cardiovascular, hepatic, coagulation, renal, and neurological systems). The SOFA score ranges from 0 to 4, with an increasing score reflecting worsening organ dysfunction, and was calculated every hour [14]. The outcome in this study was patient status in the next hour quantified by the Delta SOFA Score, which was calculated by the maximum SOFA score minus the admission SOFA score (Table 1). The seven categories of Delta SOFA score in the training set were balanced (Supplementary Table 2).

In the MIMIC-III dataset, we extracted a set of features, including demographics, vital signs, laboratory values, Elixhauser comorbidities [15], and intravenous drugs. The same drugs were combined manually (e.g. Insulin—70/30, Insulin—Glargine, Insulin—Humalog, Insulin—Humalog 75/25, Insulin—NPH, and Insulin—Regular were all incorporated into Insulin). We used the Pearson correlation coefficient [16] to select

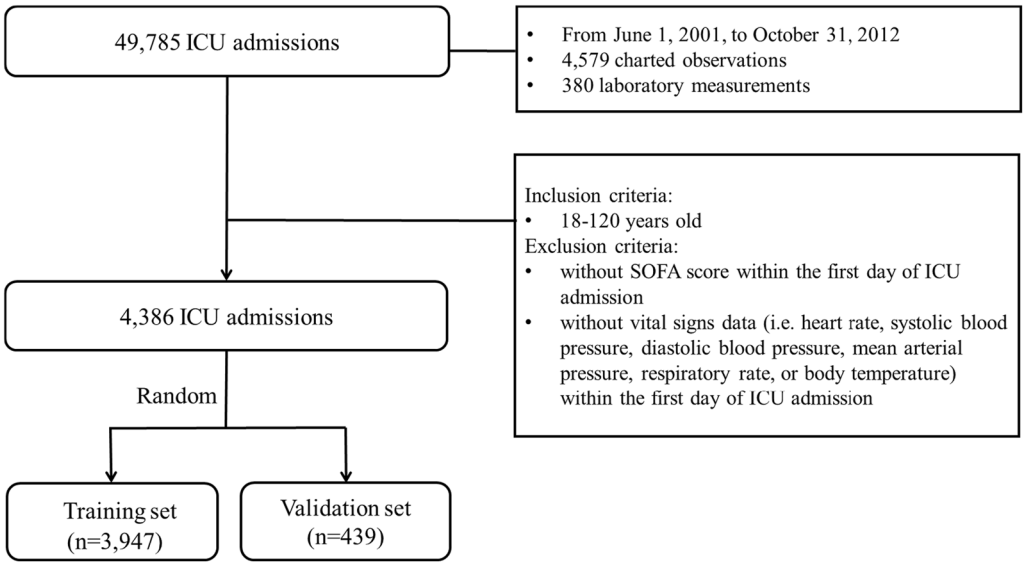


**Fig. 1** The flowchart of patient selection

**Table 1 The SOFA score definition in this study**

| | Definition |
|---|---|
| Admission SOFA score | The admission SOFA score is calculated based on the most severe value for each sub-score in the one hour after admission to ICU |
| Maximum SOFA score | The maximum SOFA score is calculated based on the most severe value for each sub-score in the current hour |
| Delta SOFA score | The delta SOFA score is calculated by the maximum SOFA score minus the admission SOFA score |
| | Categories:<br>Status 1: Delta SOFA score $\leq 0$, indicating that organ function is stable or getting better<br>Status 2: Delta SOFA score $= 1$<br>Status 3: Delta SOFA score $= 2$<br>Status 4: Delta SOFA score $= 3$<br>Status 5: Delta SOFA score $= 4$<br>Status 6: Delta SOFA score $= 5$ or 6<br>Status 7: Delta SOFA score $> 6$ |

*SOFA* sequential organ failure assessment

features relevant to the Delta SOFA scores from among the 234 features that were obtained from the MIMIC-III dataset, resulting in a final set of 186 features with a threshold greater than 0.011 (Supplementary Table 2). All features were normalized and coded as multidimensional discrete time series with 1-h time steps; those with multiple measurements within a 1-h time step were averaged (for example, heart rate). Drugs were coded as 1 or 0, where 1 indicated the use of the drug in the current hour and 0 otherwise. Drug dose and infusion rate were not of interest.

To address missing data, we exploited the forward-fill imputation strategy [17] and made further improvements as follows:

- For an ICU admission, if a feature was missing at a certain time, the nearest nonmissing value of the feature before that time was utilized.
- If all the features before the missing time were missing or the features were not measured for an ICU admission, the average value of the features across all the data were utilized.

To prevent data leakage, no information was transferred from the future to the past. In addition, to avoid data leakage between different datasets, we filled missing values in the training dataset and validation dataset separately.

The same preprocessing was performed for the eICU dataset. In addition, we excluded patients from the eICU

dataset whose feature measurement times were less than the average of those for the MIMIC-III dataset.

#### Basic clinical medical knowledge graph construction

To build a basic clinical medical knowledge graph (BCM-KG), features were mapped into the terms (i.e., entities) in the standard Unified Medical Language System (UMLS), and the related entities and relationships were retrieved. The UMLS is a compendium of health and biomedical vocabularies and standards that provides a comprehensive thesaurus and ontology of biomedical concepts and their mappings. In detail, features were mapped from continuous values to several entities representing diseases or symptoms (Supplementary Table 3); drugs were mapped to entities by names. Then, a search tool [18] was utilized to retrieve relationships and entities related to the entities linked to our 186 features with a search depth of 2. Four vocabulary sources in the UMLS were used, namely, SNOMEDCT_US [19], MDR [20], RXNORM [21], and NCI [22]. We searched 127 types of entities (Supplementary Table 4) and 976 types of relationships (as shown in the Relationship Attribute table [23]). The medical knowledge graph was finally represented by triplets, each in the form of (h, r, t), where h indicated the head entity, r was the relationship, and t represented the tail entity.

#### Basic clinical medical knowledge graph embedding

An embedding model called TransE [24] was utilized as the BCM-KG embedding model, as it can generally preserve structural information in a knowledge graph with great robustness. TransE encoded entities and relationships in the BCM-KG to a low-dimensional continuous vector space. For a triplet (*h, r, t*), the entity and relation vector were first initialized randomly. Second, a score function was defined to judge the correctness as:

$$f(h,r,t) = -h + r - t_{\frac{1}{2}},$$

which represented the negative distance between $h+r$ and $t$; higher values of $f(h,r,t)$ indicate that *(h, r, t)* had a higher probability of being true. By maximizing the score function, we finally obtained the embedding features $x^e$.

#### Prediction model

A temporal convolutional network (TCN) [25], a variant of the convolutional neural network (CNN), was used to predict the Delta SOFA score from the previous SOFA score sequence with the origin features and embedded features. We designed a status adjustment module (SAM) to capture the importance of features on the outcome under different statuses. Furthermore, there are three mechanisms that make the TCN perform well in time sequence modeling tasks: causal convolution, dilated convolution, and residual block.

**The status adjustment module (SAM)**, as shown in Fig. 2, is a weight matrix in the embedding layer. Two weight matrices are set to separately adjust the weights of the original features and embedded features in different patient states. For the original feature $x_1, x_2, \ldots, x_j, \ldots, x_n$, the SAM module adjusts them to $x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{in}$ as:

$$x_{ij} = x_i \theta_{ij},$$

where $i$ represents the current patient state, and $\theta_{ij}$ represents the weight of feature $j$ when the patient state is $i$. For the embedded feature $x_1^e, x_2^e, \ldots, x_k^e, \ldots, x_m^e$, the SAM module calculates $x_i^e$ as:

$$x_i^e = \sum_{k=1}^{m} x_k^e \varphi_{ik},$$

where $\varphi_{ik}$ represents the weight of embedded feature $k$ when the patient state is $i$, and $m$ represents the embedded feature numbers. Then, $x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{in}$ and $x_i^e$ are concatenated as the input of the TCN. The improved TCN model with SAM is called the S-TCN model in this study.

##### *Causal convolution*

The TCN uses causal convolution, in which the output of time $t$ is only convolved with elements at or before time $t$ in previous layers. The value of time $t$ in the previous layer only depends on the value of time $t$ in the next layer and its previous value. Different from that in the traditional CNN, causal convolution is a unidirectional structure, and it cannot access future data, so no information is transferred from the future to the past.

##### *Dilated convolution*

For both causal convolution and traditional convolutional neural networks; the duration of modeling is limited by the size of the convolution kernel. To obtain longer dependencies, more layers need to be stacked sequentially. TCN allows input interval sampling during convolution. Specifically, for input $\{v_0, v_1, \ldots v_i, \ldots v_{T-1}\}$ and filter $f : \{0, 1, \ldots, k-1\}$, the dilated convolution operation $F$ on the $i_{th}$ element of the input sequence is defined as:

$$F(i) = \sum_{j=0}^{k-1} f(j) \cdot v_{i-d \cdot j}$$

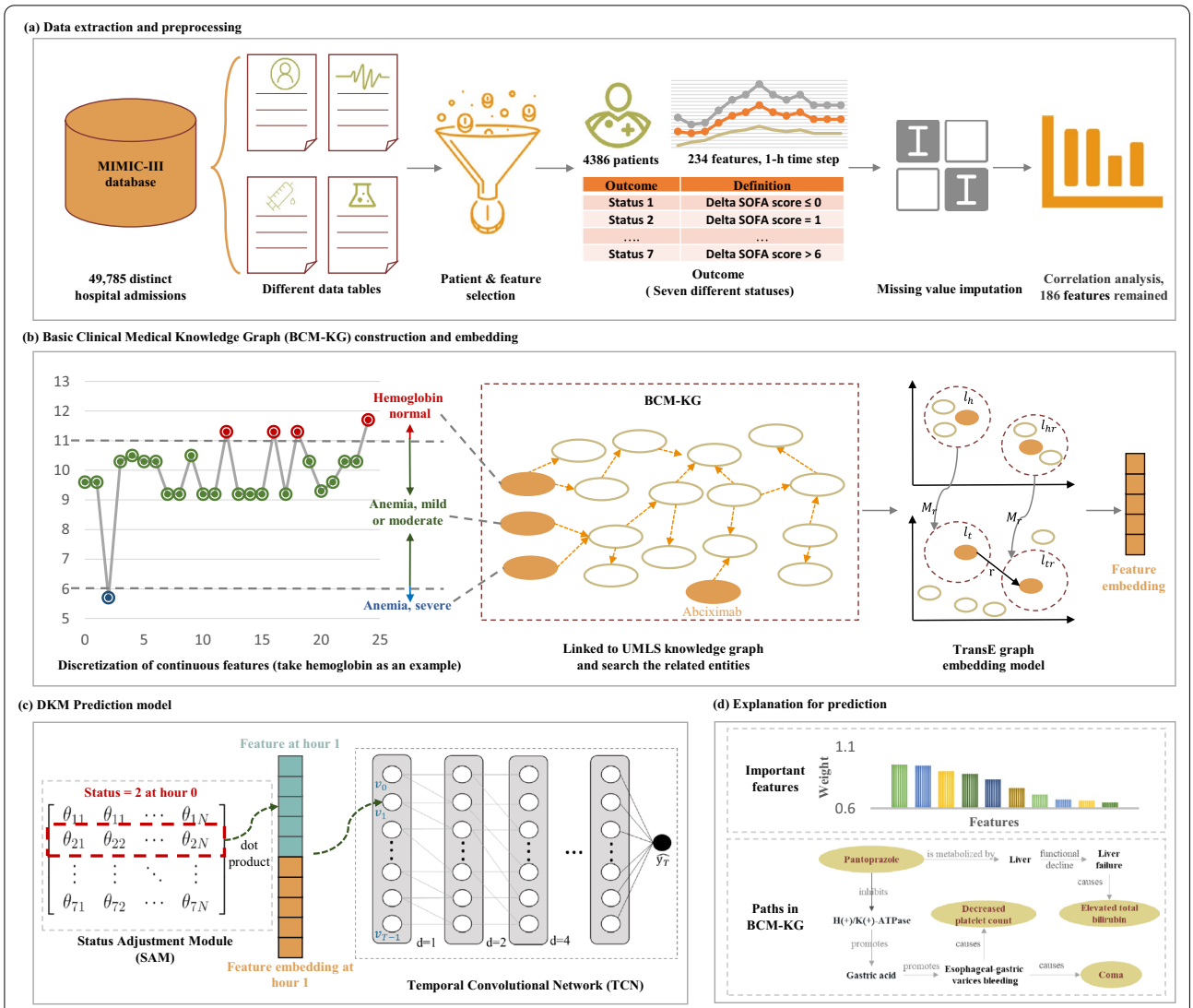


**Fig. 2** Model development overview. **a** Data extraction and preprocessing. Data on patient ICU admissions were exported from multiple tables of the MIMIC-III database and filtered according to the inclusion and exclusion criteria. A set of 186 variables were selected as the prediction model input, and the outcome was the Delta SOFA score; both took a one-hour time step. An improved forward-fill imputation strategy was used to fill missing data, and the Pearson correlation coefficient was used to select the variables that were more relevant to the outcome. **b** Basic clinical medical knowledge graph (BCM-KG) construction and embedding. Continuous variables were discretized to medical terms according to the corresponding medical standards. The search tool provided by the UMLS knowledge graph was utilized to link discretized variables to the UMLS entities and search the related entities, constructing the BCM-KG. An embedding model called TransE was utilized as the BCM-KG embedding model. **c** DKM prediction model. The status adjustment module (SAM) was added to the model embedding layer of a temporal convolutional network (TCN). After multiplying by the weight in SAM, the features were concatenated with the feature embedding and used as TCN input to predict organ failure. **d** Explanation for prediction. Patients in different organ failure statuses had different feature importance. Entities and relationships associated with important features were visualized

where $d$ is the dilation factor, $k$ is the filter size, and $i - d \cdot j$ accounts for the direction of the past. Dilation is thus equivalent to introducing a fixed step between every two adjacent filter taps. Using a larger dilation factor enables the output at the top level to represent a wider range of inputs, thus effectively expanding the receptive field with fewer layers. Given filter size $k$ and dilation factor $d$, the effective history of one such layer is *(k-1) d*. We increased d exponentially as the network deepened, i.e., $d = O(2^i)$, at level $i$ of the network, which ensures that there are some filters that hit each input within the effective history while also allowing for an extremely large, effective history using deep networks. In addition, a

spatial dropout is added after each dilated convolution for regularization.

#### *Residual connections*

A residual block [26] contains a branch leading out to a series of transformations $R$, whose outputs are added to the input $v$ of the block:

$$o = Activation(v) + R(v)$$

Since the receptive field of the TCN depends on the network depth $n$, filter size $k$, and dilation factor $d$, stabilization of a deeper and larger TCN becomes important. Each layer, more specifically, consists of multiple filters for feature extraction. Therefore, a generic residual module is employed in place of a convolutional layer.

### Model training and validation process

Data for each ICU admission were connected by a fixed separator as the model input. The parameters of the model were set as follows: the optimizer was stochastic gradient descent (SGD), the kernel size was set to 3 ($k=3$), and the remaining block level was set to 7 ($l=7$). The size of the hidden unit was 400, and the size of the embedding was set to 80. A dropout rate of 0.1 was used to avoid overfitting. The learning rate was adjustable, with an initial value of 4. When the training loss in the current epoch was greater than that in the previous three epochs, the learning rate was divided by 10. Regarding model verification, we used tenfold cross-validation to compare the model performance of DKM with that of the S-TCN and TCN on the MIMIC-III cohorts using the area under the receiver operating characteristic curve (AUC), which comprehensively evaluates the balance between sensitivity and specificity, precision, and the F1 score, which is the harmonic mean of precision and recall. The Wilcoxon signed-rank test was performed to assess the significance of the differences among these three models. In addition, we compared the performance (precision and F1 score) of the TCN model with that of three other classic deep learning models, including the gated recurrent unit (GRU), CNN and long short-term memory (LSTM). Finally, we validated our method externally on the eICU dataset using the AUC and precision. To explain the role of knowledge and the SAM, we also verified the AUC of the S-TCN and TCN in the eICU.

### Explanation for prediction

The explanation of our prediction model was described in two aspects. First, we visualized feature importance for different Delta SOFA score statuses calculated from the SAM, where $|\theta_{ij}|$ indicates the importance of the $j_{th}$ feature when the current organ failure status is $i$. Second, we presented the relations between features and other entities in the BCM-KG by exploring the paths between them using a depth-first search method [27].

## Results

### Basic statistics

The complete MIMIC-III dataset comprised 49,785 ICU admissions, corresponding to 38,597 adult patients, with information available on 4579 charted observations and 380 laboratory measurements. In this study, we included 4386 of the ICU admissions in the MIMIC-III, corresponding to 3924 patients, and 2816 admissions in the eICU. In the MIMIC-III dataset, the mean age was $56.46 \pm 13.27$ years, 58.4% of the patients were female, and the median length of ICU admission was 2 days (interquartile range 1–4). Characteristics from external validation populations were not entirely similar to those from the MIMIC-III cohort. The illness severity (initial SOFA score, Glasgow Coma Scale (GCS) scores, length of ICU stay) of the eICU patients was greater than that of MIMIC-III patients. The baseline characteristics of the two cohorts are shown in Table 2.

A total of 422,788 h of 186 features and Delta SOFA scores of the 4,386 MIMIC-III admissions were used to train and validate the prediction model, and a total of 1,048,576 h of the 2,816 eICU admissions were used to externally validate the prediction model. By retrieving relationships and entities in the UMLS related to 186 features, a BCM-KG was generated containing 101,909 entities (e.g., 'metoprolol', 'hypertension') covering all 186 features, 235 relations (e.g., 'treat', 'is a'), and 249,690 triplets (e.g., [metoprolol, treat, hypertension]).

#### *Model performance*

Using the original TCN model with the MIMIC-III cohort, the precision was 0.695, and the F1 score was 0.696. Applying the SAM to the model (i.e., S-TCN) significantly increased the precision and F1 score to 0.904 and 0.909 ($P < 0.001$), respectively. When adding the representation learning features of the BCM-KG (i.e., DKM), the DKM showed a higher precision of 0.923, F1 score of 0.927 ($P < 0.001$) and NPV of 0.989 compared with the S-TCN. The specific results of the tenfold cross-validation of the DKM are presented in Supplementary Table 5. The AUCs of these three models were 0.898, 0.966, and 0.973, respectively (Fig. 3). In addition, the TCN outperformed the other deep learning models (i.e., GRU, CNN and LSTM) (Table 3), achieving precision improvements of 0.045, 0.078 and 0.029, respectively, and F1 score improvements of 0.038, 0.149 and 0.019. Besides, the TCN outperformed the non-deep learning method logistic regression, achieving precision improvements of 0.201 and F1 score improvements of 0.189.

**Table 2 Description of MIMIC-III and eICU cohorts**

| | MIMIC-III (n = 4,386) | eICU (n = 2816) |
|---|---|---|
| Hospital characteristics | Teaching tertiary hospital | General hospitals (including 90.87% Teaching tertiary hospital) |
| Age (years) | 56.46 ± 13.27 | 56.36 ± 15.78 |
| Male gender (n, %) | 2,561, 58.4% | 1735, 61.6% |
| Premorbid complications | | |
| Hypertension (n, %) | 2,309, 52.6% | 363, 12.9% |
| Diabetes (n, %) | 357, 8.1% | 417, 14.8% |
| COPD (n, %) | 1,035, 23.6% | 239, 8.5% |
| CKD (n, %) | 746, 17.0% | 172, 6.1% |
| Congestive heart failure (n, %) | 882, 20.1% | 270, 9.6% |
| Initial SOFA score | 0 (0, 2) | 11 (9, 12) |
| Initial vital signs | | |
| Respiratory rate (/min) | 18.84 ± 5.55 | 20.56 ± 7.89 |
| Heart rate (beats/min) | 85.36 ± 17.29 | 98.40 ± 23.28 |
| Temperature site (°F) | 98.08 ± 3.44 | 96.84 ± 8.96 |
| Oxygen saturation | 89.63 ± 9.12 | 95.84 ± 5.89 |
| SBP (mmHg) | 120.52 ± 21.81 | 112.58 ± 25.53 |
| DBP (mmHg) | 64.79 ± 14.38 | 62.72 ± 18.50 |
| GCS score | 11 (11, 15) | 14 (9,15) |
| Procedures during the 24 h of data collection | | |
| Mechanical ventilation (n, %) | 1811, 41.3% | 197, 7% |
| Length of ICU stay (days) | 2.03 (1.17, 4.09) | 16.85(0.87, 110.92) |

The data are presented as the means ± standard deviations (SDs) for continuous features and as proportions for categorical features and expressed as the median interquartile range (IQR) when the data possessed a skewed distribution

*COPD* chronic obstructive pulmonary disease, *CKD* chronic kidney disease, *SOFA* sequential organ failure assessment, *SBP* systolic blood pressure, *DBP* diastolic blood pressure, *GCS* glasgow coma scale, *ICU* intensive care unit

The AUC values of the DKM, S-TCN and TCN with the eICU cohort are shown in Fig. 3. The S-TCN and TCN showed a smaller AUC than with the MIMIC-III cohort. Note that while 157 intravenous drugs were included among the input features to the training model, some were missing in the external validation set, as they were

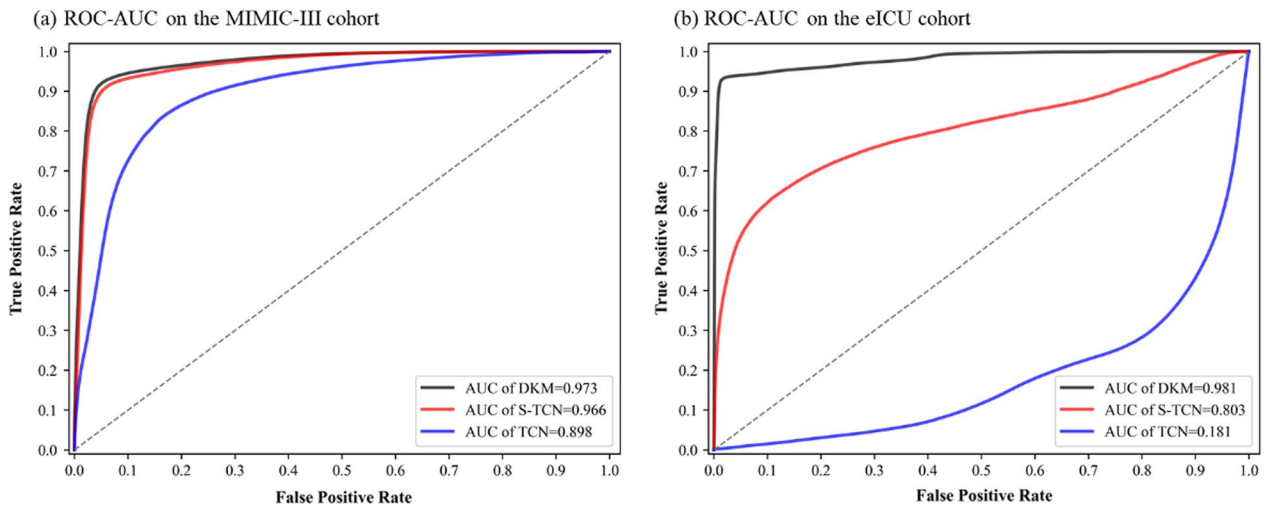


**Fig. 3** Receiver operating characteristic (ROC) curves for the organ failure prediction in MIMIC-III and eICU cohorts respectively

**Table 3 Performance of TCN and other classical deep learning models in organ failure prediction task**

| Model | Precision | F1 score |
|---|---|---|
| TCN | 0.712379 | 0.713223 |
| GRU | 0.667374 | 0.675604 |
| CNN | 0.634055 | 0.564528 |
| LSTM | 0.68311 | 0.694437 |

*TCN* temporal convolutional network, *GRU* gated recurrent unit, *CNN* convolutional neural network, *LSTM* long short-term memory

not available in the eICU (Supplementary Table 2). The final model (DKM) achieved an AUC of 0.981 (>0.75), precision of 0.860 and NPV of 0.984 on the external validation set, indicating good discriminability and calibration. Compared with the TCN and S-TCN, the final model (DKM) has a good generalization ability after introducing the knowledge graph embedded information and SAM model.

#### *Model explanation*

Feature importance can be obtained from the weight matrix of the SAM of the prediction model. Patients with different organ failure statuses had different feature importance values. If a patient is in the status of 5 Delta SOFA score, which means the organ failure is worse than admission, the top 5 important drugs for the next prediction were oseltamivir, fluconazole, amikacin, vitamin K, and nitroprusside (Fig. 4), and the top 5 important measures were the international normalized ratio, urea nitrogen level, blood platelet count, age, and hematocrit level. All feature importance values for the different organ failure states are shown in Supplementary Table 6.

We searched the hidden relations in the BCM-KG among entities with greater effects on organ failure and those representing the features included in the prediction model. In Fig. 4, each feature included in the prediction model is marked in red. We showed that knowledge can help to associate entities in organ failure prediction together through the triplets in the BCM-KG. Although metoprolol and chronic kidney disease stage 1 are not strongly relevant literally, given the linkage to BCM-KG, they can establish associations with each other through different paths.

## Discussion

Much of the contemporary focus in organ failure management is on early intervention, which allows caregivers to identify and make treatment decisions based on a reliable estimate of the deterioration of organ systems in the near future. In a well-designed randomized clinical trial, compared with standard care, the application of early warning systems was proven to help reduce the risk of unexpected events [28]. Our study extends existing real-time organ failure prediction methods, most of which collect features at a certain time granularity and predict the outcome indicators of the next period. For example, Meyer et al. [4] used 52 patient features collected every half hour to predict renal failure, and Yang et al. [5] extracted a total of 168 features collected hourly for the early detection of sepsis. In this study, the clinical data from the electronic health records of thousands of patients were constructed into a real-time prediction dataset containing 234 features collected hourly, which was used to construct a model that can predict organ failure in patients of any ICU duration in the next hour by a TCN, which has good performance (AUC of 0.973, precision of 0.923, NPV of 0.989, F1 score of 0.927) and requires less training time and memory than other sequence models.

Existing data-driven, deep learning methods usually combine their own data characteristics to make the model learn more features to achieve higher prediction accuracy. For example, the MTRNN-ATT model provided by Chen [25] utilizes the relationships between organ systems and a shared long short-term memory (LSTM) unit to exploit the correlations between different tasks for further performance improvement, achieving an accuracy of 0.899. The Classifier-GAIN model provided by Zhang [26] achieved an F1 score of 0.848 by incorporating both observed data and label information. However, the temporal characteristics of the different features used for model construction were ignored in the current models. In natural language processing, focusing on the limitation/influence of the current prediction results on the next prediction significantly improves model performance [29]. Therefore, we designed an SAM in the TCN model, capturing the influence of the current organ failure state on the next prediction. SAM can help the model capture more information from the features, that is, the different importance of features in different organ failure statuses. The experiments showed that the SAM significantly improved the prediction performance from a precision of 0.695 to 0.904.

Most existing deep learning methods do not exploit medical knowledge, leading to a bottleneck in model performance. To address these problems, approaches that incorporate prior medical knowledge and learn more model features by adding knowledge connections have been proposed to enhance the real-time predictive accuracy in this study. A knowledge graph is a data model that represents facts as nodes and relations between the nodes based on which knowledge is better structured and easier to parameterize [29]. Under a general medical KG, objects such as basic information, symptoms,

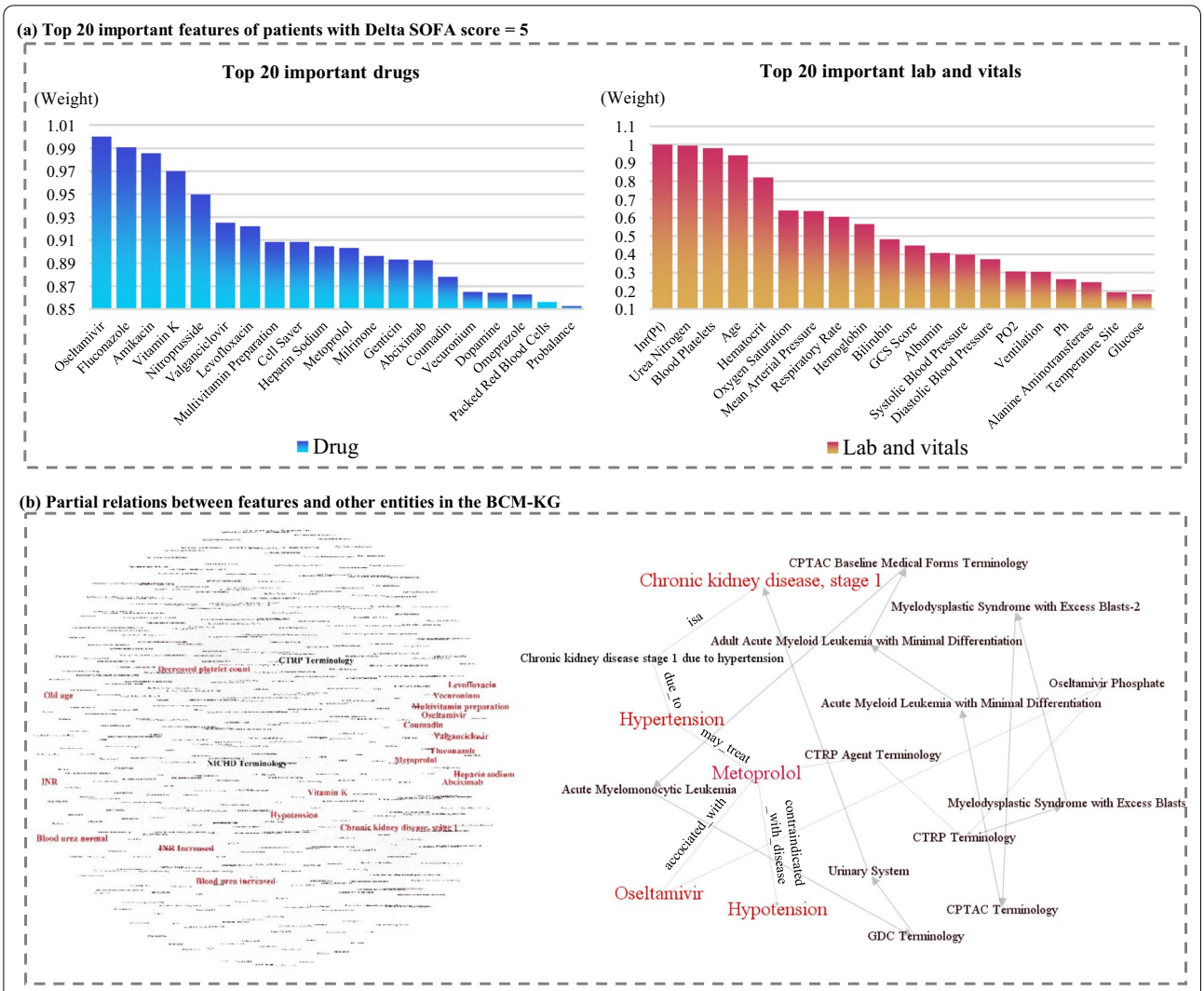


**Fig. 4** Explanation for prediction. **a** Top 20 features of patients with Delta SOFA score = 5. Variables represented by the blue columns were drugs correlated with the Delta SOFA score, and the red columns were lab and vitals correlated with the Delta SOFA score. **b** Partial relations between features and other entities in the BCM-KG. Entities used in the model were marked in red, other entities in BCM-KG were marked in black, and gray lines represented relations between entities. For example, "metoprolol" might treat "hypertension", and was contraindicated with "hypotension". *CTRP* Clinical Trials Reporting Program, *NICHD* National Institute of Child Health and Human Development, *CPTAC* Clinical Proteomic Tumor Analysis Consortium, *GDC* Genomic Data Commons

diseases, drugs, lab tests, and treatments can all be linked together through different types of referential relationships, combined with the generation of low-dimensional embedding of entities and relations, increasing feature interactions [30, 31]. Sharma et al. [32] incorporated domain knowledge via knowledge graph embedding for a UMLS, which improved the performance of the base architecture of the medical natural language inference task. To explore the interaction among the features that affect organ failure, we used the UMLS dataset to build a BCM-KG, which integrated the relations among entities in the UMLS and collected hundreds of thousands of medical knowledge elements. The representation learning features of BCM-KG improved the prediction performance from a precision of 0.904 to 0.923. This suggests that incorporating extra medical knowledge provides more features for the model, thus improving the prediction results. Furthermore, the external verification results show that our algorithm has good generalizability, with an AUC of 0.981 and precision of 0.86. The AUC from external validation without knowledge was 0.803, which was 0.178 lower than that of DKM, indicating that knowledge might be a possible mechanism for improving model performance and generalizability. Overall, models

that fused knowledge showed better performance than models based only on clinical data, which suggests that incorporating extra medical knowledge into clinical data could result in the better elucidation of clinical outcomes [33, 34].

Deep learning methods often lack explanations, which is of paramount importance for the introduction of AI models into clinical practice [35, 36]. Traditionally, there are three main types of explanation methods. The first method is the premodel explanation method, which mainly refers to preprocessing data, providing explanations from the perspective of models and data, discovering knowledge and laws from data, fully understanding the characteristics of the data distribution and understanding the problem to be solved to select the most reasonable model to approximate the optimal solution of the problem. For example, Huang et al. [37] identified the contribution of each patient feature by attention weights to improve the interpretability of the clinical predictions. Feldmann et al. [9] used the Shapley value to analyze the effect of a feature to display the feature importance on prediction. In this study, we provide pre-model interpretability by integrating the SAM in the prediction model to display the most important features contributing to the prediction in different organ failure statuses. The second method is in-model explanation methods such as linear models, parametric models or tree-based models. These models are interpretable in nature but are not suitable for the prediction of multisource time series data. The last method is the independent explanation method, which can supplement the explanatory deficiency of the black box model itself. Knowledge graphs are naturally built to be explainable [38]; many of the interactions between pathways associated with outcomes can be found in KGs as independent explanations, which plays an important role for doctors, especially primary doctors. In this study, the correlations between two indicators, chronic kidney disease (CKD) stage 1 and hypertension, which have important influences on organ failure, are shown in Fig. 4. Metoprolol may treat hypertension, and CKD may be secondary to hypertension; that is, CKD is associated with metoprolol via the hypertension pathway. Metoprolol and CKD are also linked together through a total of 11 paths, including the "Metoprolol" and "CTPR Agent Terminology" paths. We discovered the hidden relations between entities through the KG and utilized representation learning to embed components of the KG, including entities and relations, into continuous vector spaces to simplify the manipulation while preserving the inherent structure of the KG [39]. Those entity and relation embeddings can further be used to benefit all kinds of tasks, such as KG completion, entity classification, and entity resolution, which also improves the performance of our prediction model.

Our method also has limitations. First, the knowledge in BCM-KG originates from basic medical knowledge (UMLS), which is extremely broad and simple. In addition, only two-layer relations between entities were used to construct BCM-KG. These may lead to some invaluable explanations in our study. Satisfactory explanations can be achieved by introducing complete and specific professional knowledge. Additionally, this is a retrospective study, and selection bias and information bias are evitable. Although according to previous reports [40], only 6.2% of the models used prospective research data, while the remaining models used retrospective data, prospective research should be encouraged.

## Conclusion

In this study, we proposed a data-driven and knowledge-driven model to predict organ failure in critical care patients, ultimately improving the prediction accuracy and providing generalizability and independent explanations as a reference for clinical doctors. These findings support the potential of incorporating prior domain knowledge into machine learning models to inform care and service planning and provide an idea for clinically explainable research.

## Supplementary Information

The online version of this article contains supplementary material available https://doi.org/10.1007/s13755-023-00210-5.

Below is the link to the electronic supplementary material.Electronic supplementary material 1 (DOCX 117 kb)

**Funding**

The study was supported by the National Key Research and Development Program of China (No. 2022YFF0712400), the National Natural Science Foundation of China (81901842, 61906037, 62276063), Natural Science Foundation of Hunan Province (2021JJ40936), and China Primary Health Care Foundation (YLGX-WS-2020003).

## Declarations

**Conflict of interest**

None.

**Author details**

[1]School of Computer Science and Engineering, Southeast University, Nanjing 211189, People's Republic of China. [2]Department of Anesthesiology, Third Xiangya Hospital, Central South University, Changsha 410013, People's Republic of China. [3]School of Computer Science and Engineering, The University of Hong Kong, Hong Kong 999077, People's Republic of China.

Received: 10 November 2022 Accepted: 2 January 2023
Published: 23 January 2023

**References**


1. Carrico CJ, Meakins JL, Marshall J, et al. Multiple-organ-failure syndrome. Arch Surg. 1986;121(2):196–208.
2. Lambden S, Laterre PF, Levy MM, et al. The SOFA score—development, utility and challenges of accurate assessment in clinical trials. Crit Care. 2019;23(1):1–9.
3. Kim HI, Park S. Sepsis: early recognition and optimized treatment. Crit Care Med. 2019;82(1):6–14.
4. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, Lange J, Thiesson B. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun. 2020;11(1):1–11.
5. Yang M, Liu C, Wang X, et al. An explainable artificial intelligence predictor for early detection of sepsis. Crit Care Med. 2020;48(11):e1091–6.
6. Chen W, Wang S, Long G, et al. Dynamic illness severity prediction via multi-task rnns for intensive care unit. In: IEEE International Conference on Data Mining (ICDM)IEEE, pp. 917–922; 2018
7. Zhang X, Zhao Y, Callcut R, et al. Multiple Organ Failure Prediction with Classifier-Guided Generative Adversarial Imputation Networks. 2021, p^pp arXiv:2106.11878
8. Antoniadi AM, Du Y, Guendouz Y, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl Sci. 2021;11(11):5088.
9. Feldmann C, Philipps M, Bajorath J. Explainable machine learning predictions of dual-target compounds reveal characteristic structural features. Sci Rep. 2021;11(1):1–11.
10. Durán JM. Dissecting scientific explanation in AI (sXAI): a case for medicine and healthcare. Artif Intell. 2021;297:103498.
11. Rotmensch M, Halpern Y, Tlimat A, et al. Learning a health knowledge graph from electronic medical records. Sci Rep. 2017;7(1):1–11.
12. Li L, Wang P, Yan J, et al. Real-world data medical knowledge graph: construction and applications. Artif Intell Med. 2020;103:101817.
13. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. J Br Surg. 2015;102(3):148–58.
14. Vincent J-L, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intens Care Med. 1996;22:707–10.
15. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care. 2005;43(11):1130–9.
16. Benesty J, Chen J, Huang Y, et al. Pearson correlation coefficient. In: Noise reduction in speech processing
17. Lipton ZC, Kale D, Wetzel R. Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. In: Machine learning for healthcare conferencePMLR, pp. 253–270; 2016
18. McCray AT, Razi AM, Bangalore AK, et al., The UMLS Knowledge Source Server: a versatile Internet-based research tool. In: Proceedings of the AMIA Annual Fall SymposiumAmerican Medical Informatics Association, pp. 164–168; 1996
19. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. Stud Health Technol Inform. 2006;121:279.
20. Brown EG, Wood L. Wood SJDs: the medical dictionary for regulatory activities (MedDRA). Drug Saf. 1999;20(2):109–17.
21. Liu S, Ma W, Moore R, et al. RxNorm: prescription for electronic drug information exchange. IT Prof. 2005;7(5):17–23.
22. Fragoso G, de Coronado S, Haber M, et al. Overview and utilization of the NCI thesaurus. Compar Funct Genomics. 2004;5(8):648–54.
23. Abbreviations Used in Data Elements-2021AB Release. Available at: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/abbreviations.html. Accessed 12 Oct 2021
24. Bordes A, Usunier N, Garcia-Duran A, et al: Translating embeddings for modeling multi-relational data. In: Advances in neural information processing systems; 2013
25. Lea C, Vidal R, Reiter A, et al., Temporal convolutional networks: a unified approach to action segmentation. In: European Conference on Computer Vision (ECCV). Springer, Cham, pp. 47–54; 2016
26. He K, Zhang X, Ren S, et al., Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778; 2016
27. Tarjan R. Depth-first search and linear graph algorithms. SIAM J Comput. 1972;1(2):146–60.
28. Wijnberge M, Geerts BF, Hol L, et al. Effect of a machine learning–derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. JAMA. 2020;323(11):1052–60.
29. Deng S, Zhang N, Zhang W, et al., Knowledge-driven stock trend prediction and explanation via temporal convolutional network. In: Proceedings of the World Wide Web Conference (WWW), pp. 678–685; 2019
30. Wang M, Zhang J, Liu J, et al., Pdd graph: Bridging electronic medical records and biomedical knowledge graphs via entity linking. In: International Semantic Web Conference (ISWC). Springer, pp. 219–227; 2017
31. Gong F, Wang M, Wang H, et al., SMR: Medical Knowledge Graph Embedding for Safe Medicine Recommendation. 2017, p^pp arXiv:1710.05980
32. Sharma S, Santra B, Jana A, et al., Incorporating Domain Knowledge into Medical NLI using Knowledge Graphs. 2019, p^pp arXiv:1909.00160
33. Ma F, Gao J, Suo Q, et al., Risk prediction on electronic health records with prior medical knowledge. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), pp. 1910–1919; 2018
34. Bai T, Vucetic S. Improving medical code prediction from clinical text via incorporating online knowledge sources. In: Proceedings of the World Wide Web Conference (WWW), pp 72–82; 2019
35. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58.
36. Shickel B, Tighe PJ, Bihorac A, et al. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform. 2017;22(5):1589–604.
37. Chen P, Dong W, Wang J, et al. Interpretable clinical prediction via attention-based neural network. BMC Med Inform Decis Mak. 2020;20(3):1–9.
38. Rajabi E, Etminani K. Towards a knowledge graph-based explainable decision support hystem in Healthcare. Stud Health Technol Inform. 2021;281:502–3.
39. Wang Q, Mao Z, Wang B, et al. Knowledge graph embedding: a survey of approaches and applications. IEEE Trans Knowl Data Eng. 2017;29(12):2724–43.
40. Fleuren LM, Klausch TL, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intens Care Med. 2020;46(3):383–400.