# KerPrint: Local-Global Knowledge Graph Enhanced Diagnosis Prediction for Retrospective and Prospective Interpretations

**Kai Yang**[1] *, **Yongxin Xu**[3,4] *, **Peinie Zou**[3,4], **Hongxin Ding**[3,4], **Junfeng Zhao**[3,4,5] †, **Yasha Wang**[2,3,5] †, **Bing Xie**[3,4,5]

[1] Zhongguancun Laboratory, Beijing 100094, China
[2] National Engineering Research Center For Software Engineering, Peking University, Beijing 100871, China
[3] Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing 100871, China
[4] School of Computer Science, Peking University, Beijing 100871, China
[5] Peking University Information Technology Institute (Tianjin Binhai), Tianjin 300450, China
yangkai@mail.zgclab.edu.cn; {dinghx, zhaojf, wangyasha, xiebing}@pku.edu.cn; {xuyx, zplkq}@stu.pku.edu.cn

## Abstract

While recent developments of deep learning models have led to record-breaking achievements in many areas, the lack of sufficient interpretation remains a problem for many specific applications, such as the diagnosis prediction task in healthcare. The previous knowledge graph(KG) enhanced approaches mainly focus on learning clinically meaningful representations, the importance of medical concepts, and even the knowledge paths from inputs to labels. However, it is infeasible to interpret the diagnosis prediction, which needs to consider different medical concepts, various medical relationships, and the time-effectiveness of knowledge triples in different patient contexts. More importantly, the retrospective and prospective interpretations of disease processes are valuable to clinicians for the patients' confounding diseases. We propose KerPrint, a novel KG enhanced approach for retrospective and prospective interpretations to tackle these problems. Specifically, we propose a time-aware KG attention method to solve the problem of knowledge decay over time for trustworthy retrospective interpretation. We also propose a novel element-wise attention method to select candidate global knowledge using comprehensive representations from the local KG for prospective interpretation. We validate the effectiveness of our KerPrint through an extensive experimental study on a real-world dataset and a public dataset. The results show that our proposed approach not only achieves significant improvement over knowledge-enhanced methods but also gives the interpretability of diagnosis prediction in both retrospective and prospective views.

## Introduction

As the leading approach, deep learning has yielded record-breaking achievements for its ability of learning deep representations from a large volume of labeled datasets (Krizhevsky, Sutskever, and Hinton 2017; Feng et al. 2021, 2022). But in some areas where high-level model interpretability is required for data analysis tasks, such as healthcare (Choi et al. 2016; Yang et al. 2017), the application of deep learning models still faces various challenges.

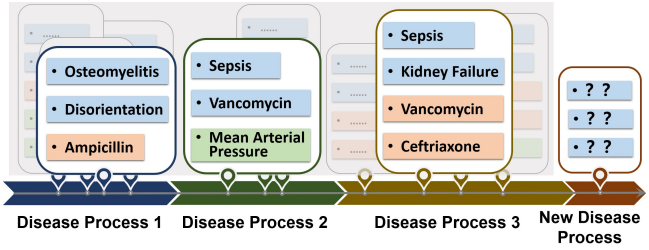


Figure 1: An example of the Sepsis process of a patient.

Diagnosis prediction is one of these typical healthcare scenarios (Choi et al. 2018; Zhang et al. 2020a). It predicts the future diagnoses of patients based on historical sequences of different clinical events (e.g., Diagnoses, Medications, Laboratory Tests and Procedures) from the Electronic Healthcare Records (EHR) Data. Recently many researchers have made progress by extracting the hierarchical semantics of various medical concepts from external resources (Yang et al. 2021), especially medical knowledge graphs(KGs), to enhance the diagnosis prediction for better performance (Chen et al. 2018) and interpretations (Ma et al. 2018b,a; Choi et al. 2017). However, there are still lots of issues to tackle with.

Some KG enhanced works try to learn clinically meaningful representations of medical concepts (Choi et al. 2017; Ma et al. 2018b) or the importance of medical concepts and historical visits (Luo et al. 2020; Zhang et al. 2020b). These works can give some knowledge-based clues instead of exact reasons or explicit evidence for the diagnosis prediction task. Other works aim to learn the importance of paths from the KG (Ye et al. 2021), as interpretable evidence from input features to a certain label specifically for a specific disease. These interpretation approaches are not feasible for the diagnosis prediction task, as the labels of diagnosis prediction are almost all disease concepts, and the candidate knowledge paths are enormous and may cover the whole KG.

Moreover, there is also a need for more consideration of the time-effectiveness of external knowledge for different

*These authors contributed equally.
†Corresponding author

scenarios. For example, *Diabetes* and *Heart Failure* have a *Complication* relationship in the KG. If the interval time between the *Diabetes* event and the *Heart Failure* event is three years or even longer, the usefulness of this knowledge triple should be diminished.

To cope with these challenges, we propose **KerPrint**, a novel **K**nowledge graph **e**nhanced approach for **r**etrospective and **Pr**ospe-ctive **int**erpretation based on a hierarchical KG attention mechanism. It includes local KGs to capture personal context-aware knowledge and global KGs to incorporate potential evolution-oriented knowledge for the diagnosis prediction task.

Our key idea is that chronic diseases generally have clear phases, from the occurrences, various developments and finally to the clinical outcomes, called the disease process. Figure 1 shows the *Sepsis* process, which mainly includes the patient's historical visit sequence and different clinical outcomes. When the patient's disease process has not changed in the to-be-predicted visit, the evolution phases, including the *Osteomyelitis* stage, the *Sepsis* stage, and the *Sepsis* with *Kidney Failure* stage, are a reasonable retrospective interpretation. However, when the patient enters a new phase of the *Sepsis* process in the next visit, new knowledge supporting the directions of disease evolution is needed to improve the model interpretation further.

Based on this observation, we propose KerPrint that combines retrospective and prospective interpretations for different phases of the disease process. Specifically, when the predictions indicate that the phase of the disease process has not changed, the KerPrint tends to review influential medical concepts and their relations utilizing a novel time-aware KG attention method, and more importantly extract relevant patterns for the retrospective interpretation. When the patient enters a new phase of the disease process, the local KG with retrospective interpretation is insufficient. Therefore, the KerPrint also integrates the global KG using a novel element-wise attention method to explore the possible evolutional direction of the disease for the prospective interpretation. This approach helps to discover the mechanism of disease evolution and to better conduct syndrome differentiation and treatments. Our contributions are listed as follows:

- We propose a novel diagnosis prediction model that can simultaneously provide retrospective and prospective interpretations respectively with local and global KGs.
- We construct a local KG for each patient that integrates the time interval information between patient visits. We also propose a novel time-aware KG attention mechanism designed to learn the context of knowledge and consider the time-effectiveness of knowledge triples.
- We propose a novel element-wise attention mechanism for deeply selecting and integrating the global KG with the patients' local KGs to explore the evolutional directions of diseases.
- We conduct extensive experiments using a real-world EHR dataset and a public MIMIC-III dataset. Experimental results and ablation studies from both datasets demonstrate that our KerPrint performs better than other knowledge-enhanced methods. The interpretation analysis also confirms the interpretability of the proposed model in both retrospective and prospective views.

## Problem Formulation

An EHR dataset is a collection of chronological records of patients' visits which include several medical codes.

**Medical Codes**. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}, c_*\}$ be the entire set of codes used in an EHR dataset, where $|\mathcal{C}|$ is the number of medical codes. Each code may represent a diagnosis, a medication, a procedure, or a laboratory test.

**EHR Dataset**. Each patient's visit sequence is defined as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$, where the $t$-th visit is denoted by a multi-hot vector $\mathbf{x}_t \in \{0,1\}^{|C|}$. The $i$-th element of one visit vector is set to 1 if it contains the medical code $c_i$.

**Medical Knowledge Graph**. The medical knowledge graph (KG) contains kinds of medical entities, such as diseases, medications, laboratory tests, procedures and their different relationships. It can be denoted in the form of triples as $\mathcal{G} = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\}$.

**Diagnosis Prediction**. Given a patient's visit sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$, corresponding time interval sequence, and medical knowledge graph $\mathcal{G}$, the goal of diagnosis prediction is to predict a binary vector $\mathbf{y}' \in \{0,1\}^{|\mathcal{L}|}$, representing the possible diagnoses in the next visit $\mathbf{x}_{T+1}$ where $\mathcal{L}$ is the number of whole diseases or disease categories.

## Methodology

The framework of the proposed KerPrint model is shown in Figure 2. It comprises three main modules:1) Local Knowledge Learning, 2) Global Knowledge Learning, 3) Prospective Knowledge Integration.

### Local KG Construction

Given a patient's visit sequence $\mathbf{X}$ and the medical KG $\mathcal{G}$, we use breadth-first search (Bundy and Wallen 1984) to generate local knowledge projections $\mathcal{G}_p = \{(c_h, r, c_t, \tau) \mid c_h, c_t \in \mathcal{E}_p, r \in \mathcal{R}_p\}$, where $\tau$ is the time interval between medical concepts $c_h$ and $c_t$ in the real visit sequence. For example, there is a relation path with time interval information: $c_1 \xrightarrow{(r_1,\tau_1)} c_3 \xrightarrow{(r_2,\tau_2)} c_8 \xrightarrow{(r_3,\tau_3)} c_9$, where $c_i, r_i$ are the corresponding entities and relations of the KG $\mathcal{G}$. The path is then split into a collection of quadruples. In this way, the local KG $\mathcal{G}_p$ is extracted, and time information is annotated for each local knowledge.

### Time-aware Attentive Embedding Propagation

Temporal information is essential for discerning the importance of associations in KGs and improving the quality of node representation learning. In order to capture non-stationary temporal changes in different diseases by embedding time into the latent visit space, we propose a novel time-aware graph attention propagation mechanism that incorporates temporal information between medical concepts into the computation of graph neural network message passing.

**Time-aware Graph Attention:** Given a quadruple $(c_h, r, c_t, \tau) \in \mathcal{G}_p$, to better model the effect of time on

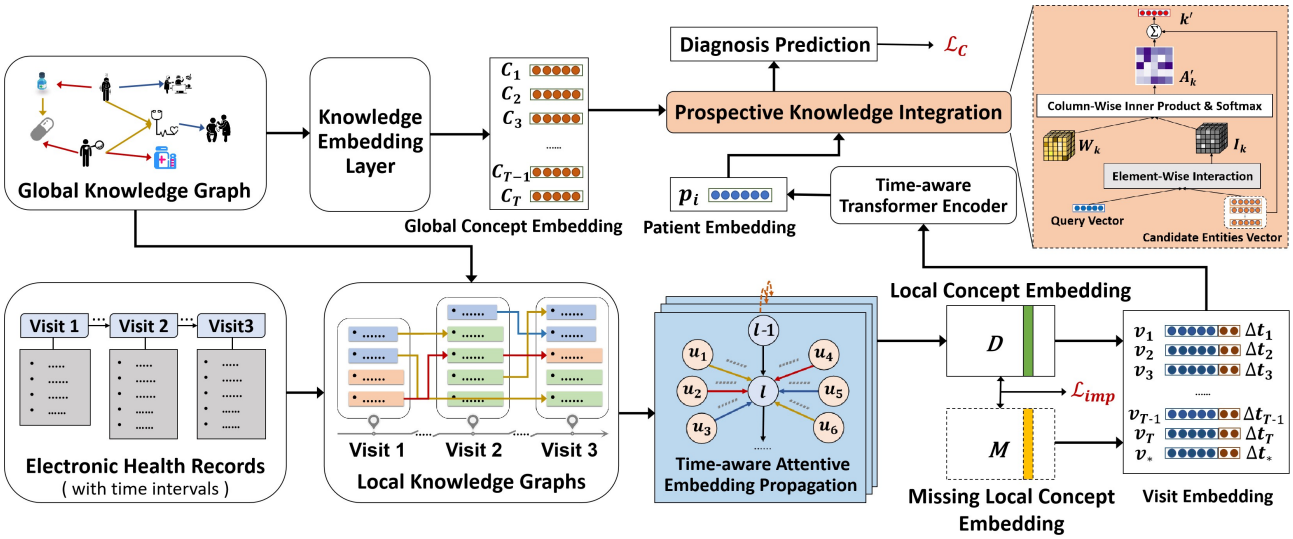


Figure 2: Illustration of the proposed KerPrint model.

the interaction of medical concepts, we first map the time interval $\tau$ into the KG entity space:

$$\mathbf{f}_\tau = \mathbf{W}_r \left( \mathbf{1} - \tanh \left( \left( \mathbf{W}_f \frac{\tau}{180} + \mathbf{b}_f \right)^2 \right) \right) + \mathbf{b}_r \quad (1)$$

where $\mathbf{W}_f \in \mathbb{R}^a$, $\mathbf{b}_f \in \mathbb{R}^a$, $\mathbf{W}_r \in \mathbb{R}^{m \times a}$ and $\mathbf{b}_r \in \mathbb{R}^m$ are all trainable parameters. Here, the shorter the interval between one medical concept and another in the patient visit sequence, the easier it is to be activated.

Given an entity $c_h$ in the local KG $\mathcal{G}_p$, we use $\mathcal{N}_h = \{(c_h, r, c_u, \tau) \mid (c_h, r, c_u, \tau) \in \mathcal{G}_p\}$ to denote the one-hop triplet collection of $c_h$. We also consider the effects of different types of relations and nodes in the KGs. The projection matrix $\mathbf{M}_r \in \mathbb{R}^{d \times k}$ is set to project entity embeddings to a relation-specific space. Then we concatenate the projected head and tail entity embeddings with the time embedding to learn the attention score $\pi(c_h, r_u^h, c_u, \tau_u^h) \in \mathbb{R}$ through a feedforward neural network $\mathbf{FFN}$:

$$\pi(c_h, r_u^h, c_u, \tau_u^h) = \mathbf{FFN}\left(\mathbf{M}_r \mathbf{e}_h \| \mathbf{M}_r \mathbf{e}_u \| \mathbf{f}_\tau\right) \quad (2)$$

The model can learn the attention weights of neighbor nodes according to the head and tail entities and the time interval, which can better solve the non-stationary performance of different nodes in KG.

Then we normalize the coefficients across all quadruples connected with $c_h$ by adopting the softmax function:

$$\widetilde{\pi}(c_h, r_u^h, c_u, \tau_u^h) = \mathbf{softmax}\left(\pi\left(c_h, r_u^h, c_u, \tau_u^h\right)\right) \quad (3)$$

**Information Propagation:** We can obtain one-hop neighbor representation of entity $c_h$ by computing the linear combination of $c_h$'s one-hop triplet collection:

$$\mathbf{e}_{\mathcal{N}_h} = \sum_{(c_h, r_u^h, c_u, \tau_u^h) \in \mathcal{N}_h} \widetilde{\pi}(c_h, r_u^h, c_u, \tau_u^h)\ \mathbf{e}_u \quad (4)$$

where $\widetilde{\pi}(c_h, r_u^h, c_u, \tau_u^h)$ indicates how much information is propagated from $c_u$ to $c_h$ conditioned to relation $r_u^h$ and time interval $\tau_u^h$, and $\mathbf{e}_u \in \mathbb{R}^d$ is the embedding for entity $c_u$.

**Information Aggregation:** We employ *GCN Aggregator* (Kipf and Welling 2016) to aggregate the entity representation $\mathbf{e}_h^{(l-1)}$ and its one-hop neighbor representation $\mathcal{N}_h^{(l-1)}$ as the new representation in the next graph network layer $l$:

$$\mathbf{e}_h^{(l)} = \text{LeakyReLU}\left(\mathbf{W}_{GCN}\left(\mathbf{e}_h^{(l-1)} + \mathbf{e}_{\mathcal{N}_h}^{(l-1)}\right)\right) \quad (5)$$

where we use LeakyReLU as the activation function. $\mathbf{W}_{GCN} \in \mathbb{R}^{d' \times d}$ are the parameters to choose useful information for propagation, and $d'$ is the transformation size.

**Medical Code Representation Learning:** After performing $L$ layers, we obtain multiple representations for entity h in KG, termed $\left\{\mathbf{e}_h^{(1)}, \cdots, \mathbf{e}_h^{(L)}\right\}$.To enrich the initial embedding, we concatenate the representations at each step into a single embedding: $\mathbf{e}_h = \mathbf{e}_h^{(0)} \| \cdots \| \mathbf{e}_h^{(L)}$.

Through analyzing associations between medical concepts by utilizing time interval information, we obtain the knowledge-matched code embedding $\mathcal{D} = \left[\mathbf{e}_{D_1}, \mathbf{e}_{D_2}, \cdots, \mathbf{e}_{D_{|\mathcal{D}|}}\right]$ which contains the information of multi-hop neighbors of each medical code. However, there are still many medical codes that cannot correspond to entities in KG, resulting in the inability to learn effectively through the higher-order connection relationships in KG. To address these issues, we propose a medical concept completion loss $\mathcal{L}_{\text{imp}}$ based on the assumption that medical concepts appearing in the same medical visit are related:

$$\mathcal{L}_{\text{imp}} = -\frac{1}{T} \sum_{t=1}^{T} \sum_{i:\mathbf{c}_i \in V_t} \sum_{j:\mathbf{c}_j \in V_t, j \neq i} \log p\left(c_j \mid c_i\right) \quad (6)$$

$$\text{where} \quad p\left(c_j \mid c_i\right) = \frac{\exp\left(\mathbf{e}_{M_j}^\top \mathbf{e}_{D_i}\right)}{\sum_{k=1}^{|\mathbf{M}|} \exp\left(\mathbf{e}_{M_k}^\top \mathbf{e}_{D_i}\right)} \quad (7)$$

where we use $\mathcal{M} = \left[\mathbf{e}_{M_1}, \mathbf{e}_{M_2}, \cdots, \mathbf{e}_{M_{|\mathcal{M}|}}\right]$ to denote the knowledge-unmatched code embedding and $|\mathcal{C}| = |\mathcal{D}| +$

$|\mathcal{M}|$. Since two medical codes are in the same visit, there is a certain correlation between them, which assists the representation learning of unmatched medical codes.

Given a patient, we then compute an embedding $\mathbf{v}_t = \frac{1}{|C_t|}\sum_{c_i \in C_t} \mathbf{e}_{c_i}$ for each visit $t$. We feed the visit vector sequence $[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_T, \mathbf{v}_*]$ into the Time-aware Transformer layer (Luo et al. 2020) for the patient's hidden state $\mathbf{h}' \in \mathbb{R}^l$.

## Global Knowledge Learning

Doctors not only focus on the patient's historical medical records but also select knowledge according to the patient's physical health status from their domain knowledge to assist prediction. Therefore, we need to improve the quality of entity representation in the global KG and simulate a robust medical domain knowledge corpus owned by doctors.

Medical KGs, e.g., SNOMED CT (Donnelly et al. 2006) contain various types of medical concepts, such as diseases, medications, lab tests, etc. In order to model better message propagation in heterogeneous medical KG, we employ a relational attention mechanism (Wang et al. 2019) to calculate the propagation weight $\pi(h, r, t)$:

$$\pi(h, r, t) = (\mathbf{M}_r \mathbf{e}_t)^\top \tanh\left((\mathbf{M}_r \mathbf{e}_h + \mathbf{e}_r)\right) \tag{8}$$

where the attention score depends on the distance between the head entity $h$ and the tail entity $t$ in the relation $r$ space. The propagation and aggregation of information here are consistent with the time-aware attention mechanism.

Then, we obtain the global knowledge embedding $\mathbf{K}_p = \left[\mathbf{e}_{k_1}, \mathbf{e}_{k_2}, \cdots, \mathbf{e}_{k_{|\mathcal{K}|}}\right]$ for each patient, where $|\mathcal{K}|$ is the number of the candidate KG nodes. We select the nodes related to the patient's historical visits but have never been present in the past visits as the candidate global knowledge.

## Prospective Knowledge Integration

To imitate the step that doctors employ their domain knowledge according to the patient's health status, we propose a novel element-wise attention module which captures the pairwise interaction between the representation of the patient and global knowledge. The details of the element-wise attention module are described in Figure 2. We first convert the patient hidden state $\mathbf{h}'$ to a query vector $\mathrm{q} \in \mathbb{R}^d$: $\mathbf{q} = \mathbf{W}_q \mathbf{h}' + \mathbf{b}_q$ ,where $\mathbf{W}_q \in \mathbb{R}^{d \times l}$ and $\mathbf{b}_q \in \mathbb{R}^d$ are trainable parameters.

Then we conduct $d \times d$ pairwise interactions of the query vector $\mathbf{q} \in \mathbb{R}^d$ and global knowledge $\mathbf{e}_{k_l} \in \mathbb{R}^d$ considering both element-wise subtraction and multiplication. The pairwise interactions of them form a matrix $\mathbf{I}_{k_l}$:

$$\mathbf{I}_{k_l}[i][j] = \mathbf{q}^{(i)} \mathbf{e}_{k_l}^{(j)} + \left(\mathbf{q}^{(i)} - \mathbf{e}_{k_i}^{(j)}\right) \tag{9}$$

where the j-th column vector of $\mathbf{I}_{k_i}$ represents the interactions between the query vector and the j-th element of the initial knowledge embedding $\mathbf{e}_{k_l}$. For each element of the initial knowledge representation $\mathbf{e}_{k_l}$, we employ a trainable weight vector $\mathbf{w}_j$ to calculate the element-wise attention score $\alpha_{k_l,j}$ of the j-th element:

$$\alpha_{k_l,j} = \mathrm{Softmax}\left(\mathbf{w}_j{}^\top \mathbf{i}_{k_1,j}, \mathbf{w}_j{}^\top \mathbf{i}_{k_2,j}, \cdots, \mathbf{w}_j{}^\top \mathbf{i}_{k_{|\mathcal{K}|},j}\right) \tag{10}$$

where $\mathbf{i}_{k_l,j}$ is the j-th column vector of $\mathbf{I}_{k_i}$. The element-wise attention score vector of $\mathbf{e}_{k_l}$ is denoted as $\alpha_{k_l} = [\alpha_{k_l,1}, \alpha_{k_l,2}, \cdots, \alpha_{k_l,d}]$. We then conduct a Hadmard product of $\alpha_{k_l}$ and initial knowledge embedding $\mathbf{e}_{k_l}$ to obtain the refined representation:

$$\mathbf{e}'_{k_l} = \alpha_{k_l} \odot \mathbf{e}_{k_l} \tag{11}$$

Thus, the final global knowledge embedding is obtained as $\mathbf{k}' = \sum_{l=1}^{|\mathcal{K}|} \mathbf{e}'_{k_l}$.

## Prediction and Optimization

We concatenate the patient hidden state $\mathbf{h}'$ and the final global knowledge embedding $\mathbf{k}'$ as the output vector $\mathbf{o} \in \mathbb{R}^{l+d}$ for the patient: $\mathbf{o} = \mathbf{h}' \oplus \mathbf{k}'$. We use a simple linear layer with a sigmoid activation function on the model output $\mathbf{o}$ to calculate the predicted label $\mathbf{y}'$. The loss function of classification for this task is the cross-entropy loss $\mathcal{L}_{\mathrm{c}}$. Finally, we combine medical concept completion loss $\mathcal{L}_{\mathrm{imp}}$, margin-based loss $\mathcal{L}_{\mathrm{TransR}}$ and cross-entropy loss $\mathcal{L}_{\mathrm{c}}$ as the final loss $\mathcal{L}(\Theta)$:

$$\mathcal{L}(\Theta) = \mathcal{L}_{\mathrm{c}} + \mathcal{L}_{\mathrm{imp}} + \lambda \|\Theta\|_2^2 \tag{12}$$

where $\Theta$ denotes all the model parameters, and we conduct $L_2$ regularization parameterized by $\lambda$ to prevent overfitting.

# Experiments

In this section, we conduct experiments on two real-world medical datasets to evaluate the performance of our proposed KerPrint method [1]. We first introduce the details of the experimental settings, including datasets, baselines, metrics, and strategies. Then we discuss the results of comparative experiments. Finally, we conduct an interpretation analysis on how KerPrint interprets the prediction results from two perspectives: *Retrospective* and *Prospective*.

## Experimental Setup

### EHR Datasets

- **MIMIC-III Dataset** is a public dataset[2] that includes various types of medical events in different visits generated by patients, such as diagnoses, lab tests, medications, etc (Johnson et al. 2016). In this study, we choose the patients who made at least 2 visits in the past 365 days and predict the diagnosis information of the next visit. We study a 135-class classification problem that predicts the diseases of a patient by grouping the ICD-9 codes into 135 higher-level medical groups using the Unified Medical Language System (Ho, Ghosh, and Sun 2014).
- **GATH Dataset** is a real-world dataset from a Grade A Tertiary Hospital in China. We utilize all the patients with at least 2 visits in their 3-year observation window. We adopt the top 150 frequent diagnoses as labels that cover 77.5% of the samples.

[1] https://github.com/xyxpku/KerPrint

[2] https://mimic.mit.edu/docs/iii/

| Dataset | Metrics | Methods w/o external knowledge | | | Methods w/ external knowledge | | | Ours |
|---|---|---|---|---|---|---|---|---|
| | | T-LSTM | Dipole | HiTANet | GRAM | KAME | $CGL_{n-}$ | KerPrint |
| MIMIC-III | MacroAUC | 0.4949(0.006) | 0.6722(0.017) | 0.7158(0.011) | 0.4678(0.007) | 0.5719(0.013) | 0.7013(0.026) | **0.7330(0.013)** |
| | MicroAUC | 0.8725(0.001) | 0.8905(0.001) | 0.9047(0.002) | 0.7942(0.003) | 0.8643(0.001) | 0.9036(0.010) | **0.9115(0.001)** |
| | MacroAUPRC | 0.0815(0.008) | 0.1491(0.007) | 0.1928(0.010) | 0.1137(0.007) | 0.1506(0.011) | 0.2023(0.010) | **0.2251(0.028)** |
| | MicroAUPRC | 0.3565(0.006) | 0.4398(0.006) | 0.5135(0.010) | 0.4119(0.008) | 0.4451(0.005) | 0.5227(0.007) | **0.5381(0.022)** |
| GATH | MacroAUC | 0.5181(0.004) | 0.6552(0.019) | 0.7375(0.011) | 0.4830(0.005) | 0.5878(0.005) | 0.7761(0.005) | **0.8223(0.011)** |
| | MicroAUC | 0.7790(0.002) | 0.8114(0.006) | 0.8225(0.002) | 0.6884(0.005) | 0.7536(0.004) | 0.8809(0.002) | **0.9013(0.006)** |
| | MacroAUPRC | 0.0285(0.007) | 0.0623(0.005) | 0.1533(0.007) | 0.1007(0.006) | 0.1089(0.009) | 0.1587(0.011) | **0.1704(0.020)** |
| | MicroAUPRC | 0.1235(0.007) | 0.1914(0.006) | 0.2633(0.006) | 0.1959(0.007) | 0.2301(0.005) | 0.3182(0.010) | **0.3273(0.015)** |

Table 1: Results for the diagnosis prediction task on MIMIC-III and GATH dataset.

| Dataset | Metrics | MedPath + HiTANet | KerPrint |
|---|---|---|---|
| MIMIC-III | AUC | 0.8190(0.005) | **0.8385(0.002)** |
| GATH | AUC | 0.7961(0.008) | **0.8438(0.012)** |

Table 2: Results for the binary classification task on MIMIC-III and GATH dataset.

| Methods | MIMIC-III | | GATH | |
|---|---|---|---|---|
| | MicroAUC | MacroAUC | MicroAUC | MacroAUC |
| KerPrint | **0.9115** | **0.7330** | **0.9013** | **0.8223** |
| $KerPrint_{m-}$ | 0.9072 | 0.7247 | 0.8787 | 0.7739 |
| $KerPrint_{t-}$ | 0.9070 | 0.7180 | 0.8931 | 0.8199 |
| $KerPrint_{k-}$ | 0.9082 | 0.7213 | 0.8955 | 0.8135 |
| $KerPrint_{a-}$ | 0.8911 | 0.7092 | 0.8785 | 0.7882 |

Table 3: Results of different KerPrint's variants.

### Medical KG Datasets

- **SNOMED CT** For the MIMIC-III dataset, we use SNOMED CT (Donnelly et al. 2006) as external knowledge, and we match medical codes to entities in the KG through mapping tables provided by SNOMED CT[3]. We filter entities and relationships related to EHR medical codes as the global KG.
- **Chinese Symptom Dataset and ICD-10 Chinese version** For the GATH dataset, we use the Chinese symptom dataset[4] and the ICD-10 Chinese version[5] as a combined external KG and complete the matchings between medical concepts to entities of KGs.

### Baseline Approaches

- **Methods without external knowledge** We use three methods without external knowledge as baselines. T-LSTM (Baytas et al. 2017) handles time intervals under a time decay assumption. Dipole (Ma et al. 2017) employs three attention mechanisms. HiTANet (Luo et al. 2020) proposes a novel time-aware Transformer with a hierarchical attention structure.
- **Methods with external knowledge** We include four methods with external knowledge as baselines. GRAM (Choi et al. 2017) is the first work to use knowledge to learn medical code representations. KAME (Ma et al. 2018b) employs external knowledge to participate in the prediction process. CGL (Lu et al. 2021) proposes collaborative graph learning to better utilize external medical knowledge. MedPath (Ye et al. 2021) extracts the personal KG and learns the disease progression information for individual patients.

Note that for CGL, we do not use the clinical notes to conduct a fair comparison, and we denote it as $CGL_{n-}$. The MedPath must find the paths from input features to the target disease entity. Thus we construct binary classification datasets for comparison with MedPath. For the MIMIC-III dataset, we set labels as the "Diseases of other endocrine glands (249–259)". For GATH Dataset, we choose "coronary artery disease, CAD" as the binary classification target. We compare with the "MedPath+HitaNet" model, which can achieve the best performance compared with other "MedPath+X" models.

**Evaluation Metrics and Strategy** For the multi-label classification task, we use micro-averaged AUROC (MicroAUC), macro-averaged AUROC (MacroAUC), micro-averaged of the Area Under the Precision-Recall Curve (MicroAUPRC) and macro-averaged AUPRC (MacroAUPRC) to evaluate the different approaches. For the binary classification task, We assess performance using the Area Under the ROC Curve (AUC). The data is randomly partitioned into three parts: training/validation/test, in a ratio of 0.70:0.15:0.15. Each model is trained with 5-fold cross-validation. Both the mean and standard deviation of test performance are reported.

## Experimental Results

Table 1 and Table 2 demonstrate the performance of the proposed KerPrint model and all the compared baselines on the two datasets. The number in () denotes the standard deviation. We observe that the KerPrint consistently achieves state-of-the-art scores on almost all metrics. Specifically, we have the following analysis:

- As methods without external knowledge, HiTANet performs well on the MIMIC-III dataset. However, the perfor-

[3]https://www.snomed.org/

[4]http://www.openkg.cn/dataset/symptom-in-chinese

[5]https://github.com/chaseliu/ICD-10-CN

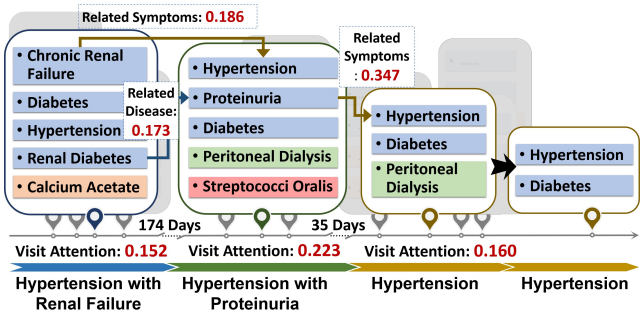


Figure 3: Retrospective Interpretations showing a reduction in *hypertensive* processes on the GATH testing set.

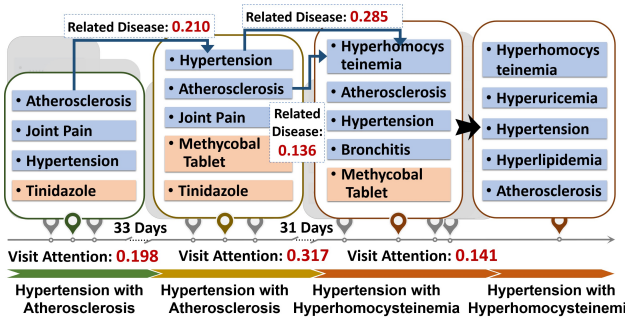


Figure 4: Retrospective Interpretations showing a increase in *hypertensive* processes on the GATH testing set.

mance of HiTANet on the GATH dataset is not outstanding. According to the statistics of these two EHR datasets, the GATH dataset is sparser than the MIMIC-III dataset, and the performance gain of KerPrint on it is larger. This illustrates the importance of introducing external knowledge to constrain the relationship between input features when they are sparse.

- Compared with other methods with external knowledge, KerPrint outperforms the above approaches. Moreover, KerPrint achieves better performance than the state-of-the-art MedPath+HiTANet. Specifically, the AUC increases by nearly 2% and 5% on the MIMIC-III dataset and the GATH dataset respectively. This demonstrates the superiority of considering the time-effectiveness of external knowledge and incorporating potential evolution-oriented knowledge to help prediction.

## Ablation Study

To evaluate the contribution of each component, we implement several variations of our framework: KerPrint without medical concept completion loss (KerPrint$_{m-}$), KerPrint without time-aware attentive embedding propagation layer (KerPrint$_{t-}$), KerPrint without global knowledge integration module (KerPrint$_{k-}$). For the sake of rigor in the medical area, we also add the KerPrint$_{a-}$ permuting the weights of time-aware attention and global knowledge integration attention (Wiegreffe and Pinter 2019).

As presented in Table 3. We observe that without medical concept completion loss, KerPrint$_{m-}$ is outperformed by KerPrint, indicating that medical concept completion loss helps to learn better representations of knowledge-unmatched medical codes and improve performance. Especially for KerPrint$_{t-}$ and KerPrint$_{k-}$, their performance drops significantly on both datasets, which indicates the importance of considering the time-effectiveness of external knowledge and integrating potential global knowledge, respectively. The performance degradation of KerPrint$_{a-}$ shows that our attention weights play a crucial role in predictions and can provide relatively faithful explanations.

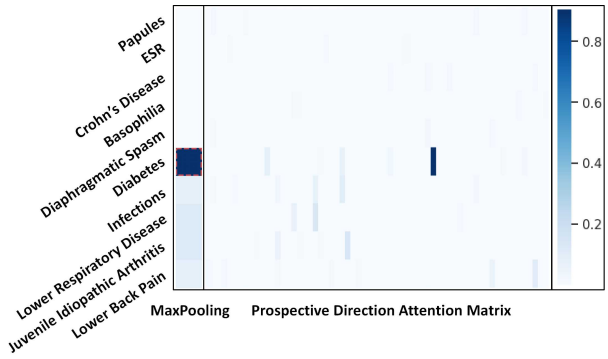


Figure 5: Prospective Interpretations showing the attention scores of candidate disease evolutional directions for a patient diagnosed with *"Type 2 Diabetes Mellitus"*.

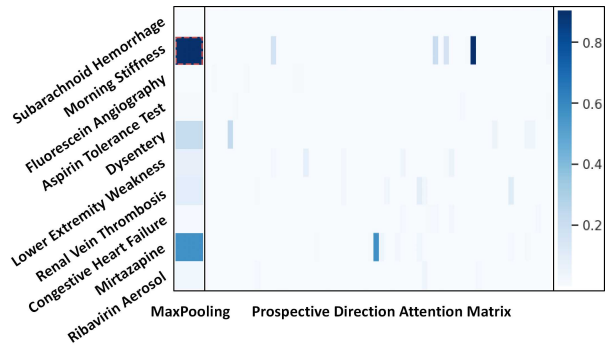


Figure 6: Prospective Interpretations showing the attention scores of candidate disease evolutional directions for a patient diagnosed with *"Severe Osteoarthritis"*.

## Model Interpretability

In this section, we analyze the explicit explanations provided by our model for the prediction results from both retrospective and prospective views through case studies. Through the integration of prospective knowledge, we can obtain whether the patient's disease process has entered a new phase. When the patient doesn't enter a new phase of the disease process, we use retrospective interpretability to interpret the prediction results. Otherwise, we use prospective interpretability.

**Retrospective Interpretations** Figure 3 shows the data of a patient diagnosed with *hypertension* and *diabetes* at the to-be-predicted visit. It describes the three visits that KerPrint

is most concerned about. Each visit is composed of many medical nodes. Nodes marked in blue, green, orange, and red represent diagnosis, treatment, medicine, and labtest events, respectively. The arrow represents the important relationship between medical nodes in the patient's local KG. The number above the arrow represents the local knowledge attention score. We can obtain that our proposed time-aware attentive propagation mechanism provides efficient information dissemination throughout these three most important visits. For example, *"Chronic Renal Failure"* diagnosed at the patient's first visit has a relationship named *"Related Symptoms"* with *"Hypertension"* diagnosed at the patient's second visit in the local KG. Even though the second visit was 174 days apart from the first visit, the attention score between them was higher than other knowledge relations, indicating that our model can capture the temporal progression regularity of different diseases based on the training data and the local KG. Through the local KG, we can also capture the patient's disease process: *Hypertension with renal failure* $\longrightarrow$ *Hypertension with proteinuria* $\longrightarrow$ *Hypertension*. We can observe that the hypertension complications of this patient are decreasing.

Figure 4 shows the data of a patient diagnosed with *Hypertension*, *Joint Pain*, *Hyperuricemia*, *Hyperlipidemia*, and *Atherosclerosis* at the to-be-predicted visit. The patient's disease process can be concluded as: *Hypertension with atherosclerosis* $\longrightarrow$ *Hypertension with atherosclerosis* $\longrightarrow$ *Hypertension with atherosclerosis and hyperhomocysteinemia*, which indicates that the patient has a progressive increase in hypertensive complications. Moreover, we can observe that at the third visit, the patient is diagnosed with a new complication "Hyperhomocysteinemia", and our proposed time-aware attentive propagation mechanism also assigns high attention scores to its relations with "Hypertension" and "Atherosclerosis", enabling the learning of medical code representations at the third visit to incorporate important information from previous visits. This is highly consistent with the medical research (Cattaneo 1999) and medical experience. When the predictions indicate that the phase of the disease process has not changed, we use the above relevant patterns to explain the results of the predictions.

**Prospective Interpretation** Figure 5 shows the data of a patient diagnosed with *Type 2 Diabetes Mellitus* (T2DM) at the to-be-predicted visit. The middle main part of the figure shows the Prospective Direction Attention Matrix calculated by the proposed element-wise attention mechanism. The left column of the figure shows the result of row-wise max pooling on the attention matrix. The darker boxes mean that the patient's query embedding is more concerned with knowledge, and vice versa. We can observe that KerPrint is able to assign the highest attention weight to *"Diabetes"* which is the broader concept of T2DM based on the patient's health status, even if he hadn't been diagnosed with T2DM during his previous visits.

Figure 6 shows the data of a patient diagnosed with *Severe Osteoarthritis* at the to-be-predicted visit. We can observe that KerPrint pays more attention to *"Morning Stiffness"*. According to medical research (Zhang et al. 2010), *Morning Stiffness* is the key symptom for the diagnosis of *Osteoarthritis*. This also shows the reasonableness of KerPrint with reliable explanations on prediction results. For these patients who go into a new phase of the disease process, even if the label has not appeared in their previous visits, we select the global knowledge related to their historical visits as the possible evolutional direction of the disease based on the learned comprehensive representation.

## Related Work

Recently, deep learning techniques have shown excellent performance in disease risk prediction (Xu et al. 2022), medication recommendation (An et al. 2021), and diagnosis prediction (Chen et al. 2018). We focus on diagnosis prediction studies from two following perspectives.

**Models focusing on capturing sequence correlations** These works focus on capturing sequential features or contextual dependencies between patient visit sequences. For example, RETAIN (Choi et al. 2016) proposes a two-level attention mechanism based on RNN, which can learn the weights of visit levels and medical code levels. Dipole (Ma et al. 2017) uses bidirectional RNNs to model EHR data and employs three attention mechanisms to calculate the correlations between medical visits. SAnD (Song et al. 2018), LSAN (Ye et al. 2020) and HiTANet (Luo et al. 2020) employ Transformer as the backbone. In addition, T-LSTM (Baytas et al. 2017), Timeline (Bai et al. 2018), Concare (Ma et al. 2020), and HiTANet (Luo et al. 2020) consider the time interval between consecutive visits.

**Models focusing on incorporating external knowledge** These works focus on leveraging external medical KGs to improve representation learning, thereby improving prediction performance. For example, some studies employ ancestor information of nodes in the KG to improve the quality of medical representation learning, such as GRAM (Choi et al. 2017), KAME (Ma et al. 2018b). Other studies employ graph neural networks (GNN) to capture high-order connection information in KGs and fuse them into the representation of inputs, such as MedPath (Ye et al. 2021), CGL (Lu et al. 2021). However, these works merely give some knowledge-based clues instead of explicit reasons or evidences for the diagnosis prediction. They also lack consideration of the impact of time information in different contexts of knowledge use.

## Conclusions

Diagnosis prediction is one of the critical tasks in healthcare. To better integrate knowledge with contexts, we propose a novel diagnosis prediction framework, named KerPrint. It employs a time-aware KG attention approach to learn the context of knowledge. We also propose an element-wise attention method to deeply select and integrate the global KG according to the patients' medical representations learned from the local KG. The effectiveness of the proposed KerPrint is verified with extensive experiments on two EHR datasets. Besides, KerPrint also provides the model interpretability in both retrospective and prospective views.

## Acknowledgments

This research is supported by the National Natural Science Foundation of China (No.62102007), the National Natural Science Foundation of China (No.62172011) and the Fundamental Research Funds for the Central Universities.

## References

An, Y.; Zhang, L.; You, M.; Tian, X.; Jin, B.; and Wei, X. 2021. MeSIN: Multilevel selective and interactive network for medication recommendation. *Knowledge-Based Systems*, 233: 107534.

Bai, T.; Zhang, S.; Egleston, B. L.; and Vucetic, S. 2018. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 43–51.

Baytas, I. M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A. K.; and Zhou, J. 2017. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 65–74.

Bundy, A.; and Wallen, L. 1984. Breadth-first search. In *Catalogue of artificial intelligence tools*, 13–13. Springer.

Cattaneo, M. 1999. Hyperhomocysteinemia, atherosclerosis and thrombosis. *Thrombosis and haemostasis*, 81: 165–176.

Chen, J.; Li, K.; Rong, H.; Bilal, K.; Yang, N.; and Li, K. 2018. A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. *Information Sciences*, 435: 124–149.

Choi, E.; Bahadori, M. T.; Song, L.; Stewart, W. F.; and Sun, J. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 787–795.

Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; and Stewart, W. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29.

Choi, E.; Xiao, C.; Stewart, W.; and Sun, J. 2018. MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Donnelly, K.; et al. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121: 279.

Feng, Y.; Wang, J.; Wang, Y.; and Chu, X. 2022. Towards Sustainable Compressive Population Health: A GAN-based Year-By-Year Imputation Method. *ACM Transactions on Computing for Healthcare*.

Feng, Y.; Wang, J.; Wang, Y.; and Helal, S. 2021. Completing Missing Prevalence Rates for Multiple Chronic Diseases by Jointly Leveraging Both Intra-and Inter-Disease Population Health Data Correlations. In *Proceedings of the Web Conference 2021*, 183–193.

Ho, J. C.; Ghosh, J.; and Sun, J. 2014. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 115–124.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM*, 60(6): 84–90.

Lu, C.; Reddy, C. K.; Chakraborty, P.; Kleinberg, S.; and Ning, Y. 2021. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. *arXiv preprint arXiv:2105.07542*.

Luo, J.; Ye, M.; Xiao, C.; and Ma, F. 2020. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 647–656.

Ma, F.; Chitta, R.; Zhou, J.; You, Q.; Sun, T.; and Gao, J. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1903–1911.

Ma, F.; Gao, J.; Suo, Q.; You, Q.; Zhou, J.; and Zhang, A. 2018a. Risk Prediction on Electronic Health Records with Prior Medical Knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, 1910–1919. New York, NY, USA: Association for Computing Machinery. ISBN 9781450355520.

Ma, F.; You, Q.; Xiao, H.; Chitta, R.; Zhou, J.; and Gao, J. 2018b. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 743–752.

Ma, L.; Zhang, C.; Wang, Y.; Ruan, W.; Wang, J.; Tang, W.; Ma, X.; Gao, X.; and Gao, J. 2020. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 833–840.

Song, H.; Rajan, D.; Thiagarajan, J. J.; and Spanias, A. 2018. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-second AAAI conference on artificial intelligence*.

Wang, X.; He, X.; Cao, Y.; Liu, M.; and Chua, T.-S. 2019. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 950–958.

Wiegreffe, S.; and Pinter, Y. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Xu, Y.; Ying, H.; Qian, S.; Zhuang, F.; Zhang, X.; Wang, D.; Wu, J.; and Xiong, H. 2022. Time-aware Context-Gated Graph Attention Network for Clinical Risk Prediction. *IEEE Transactions on Knowledge and Data Engineering*.

Yang, K.; Li, X.; Liu, H.; Mei, J.; Xie, G.; Zhao, J.; Xie, B.; and Wang, F. 2017. TaGiTeD: Predictive task guided tensor decomposition for representation learning from electronic health records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Yang, K.; Luo, Z.; Gao, J.; Zhao, J.; Ooi, B. C.; and Xie, B. 2021. LDA-Reg: Knowledge Driven Regularization using External Corpora. *IEEE Transactions on Knowledge and Data Engineering*.

Ye, M.; Cui, S.; Wang, Y.; Luo, J.; Xiao, C.; and Ma, F. 2021. Medpath: Augmenting health risk prediction via medical knowledge paths. In *Proceedings of the Web Conference 2021*, 1397–1409.

Ye, M.; Luo, J.; Xiao, C.; and Ma, F. 2020. Lsan: Modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1753–1762.

Zhang, M.; King, C. R.; Avidan, M.; and Chen, Y. 2020a. *Hierarchical Attention Propagation for Healthcare Representation Learning*, 249–256. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379984.

Zhang, W.; Doherty, M.; Peat, G.; Bierma-Zeinstra, M.; Arden, N.; Bresnihan, B.; Herrero-Beaumont, G.; Kirschner, S.; Leeb, B.; Lohmander, L.; et al. 2010. EULAR evidence-based recommendations for the diagnosis of knee osteoarthritis. *Annals of the rheumatic diseases*, 69(3): 483–489.

Zhang, X.; Qian, B.; Cao, S.; Li, Y.; Chen, H.; Zheng, Y.; and Davidson, I. 2020b. *INPREM: An Interpretable and Trustworthy Predictive Model for Healthcare*, 450–460. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379984.