MedPath: Augmenting Health Risk Prediction via Medical Knowledge Paths

Muchao Ye*
The Pennsylvania State University
USA
muchao@psu.edu

Suhan Cui*
Northeastern University
USA
suhan.cui@gmail.com

Yaqing Wang
Purdue University
USA
wang5075@purdue.edu

Junyu Luo
The Pennsylvania State University
USA
junyu@psu.edu

Cao Xiao
IQVIA
USA
cao.xiao@iqvia.com

Fenglong Ma†
The Pennsylvania State University
USA
fenglong@psu.edu

ABSTRACT

The broad adoption of electronic health records (EHR) data and the availability of biomedical knowledge graphs (KGs) on the web have provided clinicians and researchers unprecedented resources and opportunities for conducting health risk predictions to improve healthcare quality and medical resource allocation. Existing methods have focused on improving the EHR feature representations using attention mechanisms, time-aware models, or external knowledge. However, they ignore the importance of using personalized information to make predictions. Besides, the reliability of their prediction interpretations needs to be improved since their interpretable attention scores are not explicitly reasoned from disease progression paths. In this paper, we propose `MedPath` to solve these challenges and augment existing risk prediction models with the ability to use personalized information and provide reliable interpretations inferring from disease progression paths. Firstly, `MedPath` extracts personalized knowledge graphs (PKGs) containing all possible disease progression paths from observed symptoms to target diseases from a large-scale online medical knowledge graph. Next, to augment existing EHR encoders for achieving better predictions, `MedPath` learns a PKG embedding by conducting multi-hop message passing from symptom nodes to target disease nodes through a graph neural network encoder. Since `MedPath` reasons disease progression by paths existing in PKGs, it can provide explicit explanations for the prediction by pointing out how observed symptoms can finally lead to target diseases. Experimental results on three real-world medical datasets show that `MedPath` is effective in improving the performance of eight state-of-the-art methods with higher F1 scores and AUCs. Our case study also demonstrates that `MedPath` can greatly improve the explicitness of the risk prediction interpretation.[1]

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Applied computing** → **Health informatics**.

KEYWORDS

risk prediction; model interpretability; graph neural network; healthcare informatics

ACM Reference Format:
Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. MedPath: Augmenting Health Risk Prediction via Medical Knowledge Paths. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3442381.3449860

1 INTRODUCTION

With the larger accumulation and wider availability of electronic health record (EHR) data, clinicians and researchers are allowed to conduct health risk predictions to improve healthcare quality and medical resource allocation. Health risk prediction refers to using historical EHR data of patients to predict whether they will suffer from a target disease in a near future. However, since EHR data are usually encoded by high dimensional medical codes such as ICD codes[2], it is still challenging to extract meaningful signals from EHRs for risk prediction task.

To tackle this challenge, deep learning techniques have been used and demonstrated state-of-the-art performance in modeling high dimensional EHR data for health risk prediction [1, 2, 4, 6, 14, 16, 17, 21]. However, EHR data are usually sparse, and only a part of medical codes can be learned sufficiently, which limits these approaches to greatly unleash the prediction power. To address this issue, researchers propose to incorporate prior medical knowledge or knowledge graph (KG) on the web to enhance the representation learning of medical codes, which further augments health risk prediction task [5, 18–20, 30]. For example, GRAM [5] learns the latent embedding of a clinical code (e.g., diagnosis code) as a convex

* indicates equal contribution. This work was done when Suhan Cui remotely worked with Fenglong Ma.
† indicates corresponding author.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449860

[1] The code is publicly available at https://github.com/machinelearning4health/MedPath.
[2] https://www.cdc.gov/nchs/icd/

combination of the embeddings of the code itself and its ancestors on the ontology graph. KAME [20] is built upon GRAM and tries to use high-level knowledge to further improve the performance. DG-RNN [30] introduces a dynamic attention mechanism to enhance the embedding learning of medical codes with a medical KG named KnowLife [9]. Although these approaches are effective for risk prediction task, they still suffer from the following issues.

Necessity of Incorporating Personalized Knowledge Graph. All the aforementioned approaches need to encode the entire ontology or knowledge graph, which contains a large number of medical codes and corresponding relations. However, the number of overlapping medical codes between individual patients' EHR data and the entire KG is very small. Thus, using the whole KG for individual patient prediction may introduce noise information and further hurt the performance. Moreover, the leading causes of a specific target disease for different patients vary a lot, which indicates the necessity of inferring disease causes for different patients individually. Therefore, it is essential and reasonable to extract personalized knowledge graph for each patient and harness it with deep learning models for achieving personalized prediction.

Explicit Reasoning over Disease Progression Paths. Existing approaches aim to incorporate medical KGs to enhance the representation learning of medical codes with the help of relations among them and conduct *implicit reasoning* on the prediction results. However, there exist *explicit disease progression paths in KG* from the observed symptoms (i.e., medical codes) to the target disease, which are ignored by existing studies. In addition, existing approaches often exploit attention weights to identify important medical codes for the predictions, which can be considered as *implicit one-hop paths*, i.e., from the selected codes to the target disease. However, some one-hop paths may not be in accord with current medical knowledge, which leads to the untrustworthiness of existing models of both patients and doctors. Therefore, explicit reasoning and reliable explanation generation are equally important and challenging for health risk prediction task.

Our Approach. To address the aforementioned challenges, we propose a model-agnostic and ready-to-use framework, named `MedPath` as shown in Figure 1(a), which enables any existing risk prediction models to provide personalized prediction and explicit reasoning. During the **prediction** stage, `MedPath` takes both medical codes in a patient's EHR data and the target disease codes as inputs to extract a personalized knowledge graph (PKG) from a large, complex, and noisy medical knowledge graph for each patient. With the guidance of the patient representation obtained from EHR data by existing risk prediction approach, `MedPath` then conducts explicit reasoning via learning personalized knowledge representations over the extracted PKG. Finally, `MedPath` makes predictions based on both patient representations and personalized knowledge representations. During the **reasoning** stage, `MedPath` finds all possible paths within the extracted graph linking the nodes of observed symptoms and the target disease, and uses the learned weights of different paths to explicitly explain the prediction results.

Contributions. Our main contributions are listed as follows:

- We propose a general, model-agnostic and ready-to-use framework `MedPath`, which enables any existing risk prediction models to work with both EHR data and personalized knowledge graphs simultaneously, for providing personalized prediction and explicit reasoning.
- We propose a novel graph attention network to calculate weights for different paths within the extracted personalized knowledge graph, which allows `MedPath` to provide explicit, reliable, and trustworthy medical knowledge path-based explanations for the predictions.
- We conduct experiments on three real-world medical claim datasets to demonstrate the effectiveness of `MedPath` framework in improving the risk prediction performance of eight state-of-the-art risk prediction models. We also use case studies to demonstrate that `MedPath` can provide reasonable explanation for risk prediction.

2 DATA & TASK

In this section, we will introduce the format of electronic health records (EHR) and medical knowledge graph data. In addition, we will present how to preprocess those data as the inputs and then formally define our task.

2.1 Electronic Health Records

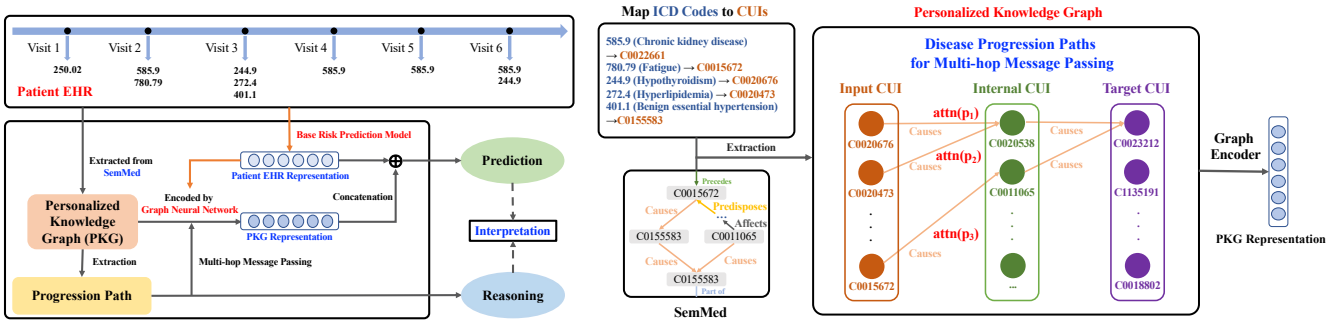
EHR is a special kind of data that is extensively used in health risk prediction task, which contains the complete medical history of a patient and provides rich information on their future status. The EHR $\mathbf{X}$ of a specific patient normally consists of records of multiple visits $[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$ where T is the number of visits, and each visit $\mathbf{x}_t$ $(1 \leq t \leq T)$ is described by several medical (ICD-9[3]) codes where each code represents a symptom, an abnormal finding, or a treatment.

The goal of this work is to predict the probability of a patient suffering from a given target disease in the future by analyzing the historical EHR data $\mathbf{X}$. Though this is a binary classification task, a specific target disease can be denoted by a set of ICD-9 codes due to its various types. Take the heart failure disease as an example. "428.0" represents *congestive heart failure*,[4] "428.1" denotes *left heart failure*,[5] and "428.2" means *systolic heart failure*.[6] For a specific target disease, we let $\mathcal{Y}$ denote the set of the target disease codes. Note that there is no overlapping code between $\mathbf{X}$ and $\mathcal{Y}$.

2.2 Medical Knowledge Graph

As mentioned before, our method relies on an external knowledge graph for explicit interpretability. The source of the medical knowledge graph that we use is SemMed [12, 22],[7] which is a huge multi-relational medical knowledge graph with more than 150,000 entities and 64 relation types. The knowledge in SemMed is described in the form of triples, which are extracted from abstract sentences of medical publications on PubMed.[8] Each triple consists of three elements: the **head entity**, the **tail entity**, and the **relation**. The head entity is the subject concept, the tail entity is the object concept, and the relation describes the relationship between the subject concept and the object concept. For example, *<Hypertensive disease, CAUSES,*

[3] https://www.cdc.gov/nchs/icd/icd9.htm
[4] http://www.icd9data.com/2015/Volume1/390-459/420-429/428/428.0.htm
[5] http://www.icd9data.com/2015/Volume1/390-459/420-429/428/428.1.htm
[6] http://www.icd9data.com/2015/Volume1/390-459/420-429/428/428.2.htm
[7] https://skr3.nlm.nih.gov/SemMed/
[8] https://pubmed.ncbi.nlm.nih.gov/



(a) Framework (b) PKG Processing

Figure 1: (a) The framework of MedPath, which augments risk prediction models by learning a PKG embedding extracted from SemMed for prediction. Disease progress can be reasoned from PKG in (b), which is helpful for explicit interpretation.

Left heart failure> is a triple in SemMed. *Hypertensive disease* is the **head entity**, *Left heart failure* is the **tail entity**, and *CAUSES* is the **relation**. Such triples not only augment the risk prediction feature learning but can also be used to explicitly interpret the main reason why the patient will suffer the target disease in the future.

2.3 Personalized Graph Extraction

Intuitively, the leading causes of a specific target disease for different patients vary a lot. Thus, it is essential to infer such reasons for different patients individually. To achieve this goal, since we are given a knowledge graph containing giant medical knowledge, we can extract a personalized knowledge graph $\mathcal{G}$ from the whole SemMed based on their EHR. The process of obtaining $\mathcal{G}$ with patient EHR data $\mathbf{X}$ and the target disease code set $\mathcal{Y}$ are in the following three steps.

2.3.1 Unification of ICD Codes and SemMed Entities. To get the personalized graph, the first problem we need to consider is that SemMed uses Concept Unique Identifiers (CUIs) to represent entities,[9] while EHR data uses ICD codes. To unify them, we first map the ICD codes to CUIs using SNOMED CT [8] for convertion. In particular, all the unique codes in $\mathbf{X}$ are mapped to a new CUI set $\mathcal{E}_x$. Also, the target code set $\mathcal{Y}$ is mapped to another CUI set $\mathcal{E}_y$. Note that an ICD code may have multiple corresponding CUIs. In this case, we map such ICD codes to the first CUIs, which have the highest matching scores. Besides, there are some ICD codes not having corresponding CUIs, we then ignore them.

2.3.2 Path Extraction. To construct the personalized graph for reasoning the causes of disease, we need to extract paths between the input CUIs $\mathcal{E}_x$ and the target CUIs $\mathcal{E}_y$. Specifically, given a CUI entity $e_x \in \mathcal{E}_x$ and a target CUI $e_y \in \mathcal{E}_y$, we use depth-first search [25] to generate all possible paths with multiple hops that link e_x and e_y. There are 64 types of relations in SemMed, but only a few of them are related to disease progression. Here, we only keep the paths with the following nine relations in the set of $\mathcal{R}$, namely *AFFECTS*, *AUGMENTS*, *CAUSES*, *DIAGNOSES*, *INTERACTS_WITH*, *PART_OF*, *PRECEDES*, *PREDISPOSES*, and *PRODUCES*. On the generated paths from $\mathcal{E}_x$ to $\mathcal{E}_y$, there are several internal CUIs on the paths, and the set of the internal CUI nodes are represented by $\mathcal{E}_i$.

[9] https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_005.html

Algorithm 1: MedPath for Health Risk Prediction

Inputs: EHR input $\mathbf{X}$, Personalized Knowledge Graph $\mathcal{G}$
Output: Predicted label y', medical paths $\{p_1, ..., p_{k^*}\}$
Encode EHR into feature $\mathbf{s}$ by Eq. (1)
Update node embeddings in $\mathcal{G}$ by Eq. (2), (4), (5) and (6) and obtain $[\mathbf{h}'_1, \cdots, \mathbf{h}'_n]$
Output predicted label y' by following Section 3.3 given $\mathbf{s}$ and $[\mathbf{h}'_1, \cdots, \mathbf{h}'_n]$
Output medical paths by following Section 3.4

2.3.3 Personalized Graph. Using the outputs from the previous two steps, we can finally obtain the personalized knowledge graph $\mathcal{G} = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\}$ with each triple (h, r, t) describing a relation r between entity h and entity t, where $\mathcal{E} = \{\mathcal{E}_x, \mathcal{E}_i, \mathcal{E}_y\}$, and $\mathcal{R}$ is the set of the selected relations.

2.4 Health Risk Prediction Task

The proposed method for the risk prediction task is formulated as follows. For each patient, given the EHR $\mathbf{X}$ and the extracted personalized knowledge graph $\mathcal{G}$, we aim to predict whether the patient would suffer from the target disease in the future. In the meantime, our method also decodes paths from the knowledge graph $\mathcal{G}$ that starts from an input EHR CUI (i.e., $e_x \in \mathcal{E}_x$) and end in a target disease CUI (i.e., $e_y \in \mathcal{E}_y$) to help explain the possible disease development process.

3 METHODOLOGY

As Figure 1 shows, MedPath is a flexible framework that can be built on any existing risk prediction model F_e to provide explicit interpretation for the prediction process via the graph neural network F_g. The first part F_e uses state-of-the-art deep learning models as the EHR encoder, which takes the EHR data $\mathbf{X}$ as the input and encode $\mathbf{X}$ into a representation vector $\mathbf{s}$. Then for the graph encoder F_g, which is a multi-relational graph neural network, it takes $\mathbf{s}$ and the extracted personalized knowledge graph $\mathcal{G}$ as inputs to perform

multi-relational reasoning. With these modules working together, the proposed MedPath framework incorporates external knowledge $\mathcal{G}$ to augment risk prediction representation $\mathbf{s}$ for final prediction. Moreover, we can have explicit interpretability from $\mathcal{G}$ on the relations between symptoms $\mathbf{X}$ and target diseases $\mathcal{Y}$ when we need to explain how MedPath makes prediction. The aforementioned process in pseudocode is shown in Algorithm 1.

3.1 EHR Encoder

As mentioned before, MedPath is a general framework which can use any existing risk prediction model as its EHR encoder to represent the data $\mathbf{X}$. A number of existing methods [1, 2, 4, 6, 14, 16, 17, 21] do not exploit external knowledge for risk prediction. However, they are still good at learning a comprehensive representation of $\mathbf{X}$ since they adopt effective and state-of-the-art deep learning techniques like recurrent neural networks and attention mechanisms. Considering that these representations are good for the training of graph encoder F_g and the final prediction, we simply remove the final classification layer of any given risk prediction model and use it as the EHR encoder F_e. So we have

$$\mathbf{s} = F_e(\mathbf{X}), \tag{1}$$

where $\mathbf{s} \in \mathbb{R}^{d_s}$ is the representation vector learned from EHR encoder, and d_s is the dimension size of $\mathbf{s}$. Feature $\mathbf{s}$ contains the abstract information of EHR extracted by advanced deep learning models, and it can aid the training of graph encoder later on.

3.2 Graph Encoder

Using the personalized graph $\mathcal{G}$ extracted for the EHR $\mathbf{X}$ in Section 2.3, we introduce the graph encoder module F_g to synthesize the graph knowledge $\mathcal{G}$ into a comprehensive vector $\mathbf{g}$. Suppose that the personalized graph $\mathcal{G}$ has n entities (nodes), the embedding of each entity $\mathbf{h}_j \in \mathbb{R}^{d_h}$ $(1 \le j \le n)$ is initialized by pre-training triples in $\mathcal{G}$ with the TranE [3] algorithm. We will conduct type-specific transformation and multi-hop message passing to embed the correlation with other entities to learn a final entity embedding $\mathbf{h}'_j$ for each entity e_j. Eventually, we get the graph representation $\mathbf{g}$ by performing the attentive pooling over all the target disease entity features $\{\mathbf{h}'_j | e_j \in \mathcal{E}_y\}$.

3.2.1 Type-Specific Transformation. There are three types of entities in the personalized knowledge graph $\mathcal{G}$: input CUIs $\mathcal{E}_x$, target CUIs $\mathcal{E}_y$, and internal CUIs $\mathcal{E}_i$. To learn a good personalized graph embedding, it is helpful to distinguish the types of different entities and embed the node type information into the node embeddings. Therefore, we apply type-specific transformation on the initialized node embeddings to embed the node type information,

$$\mathbf{v}_j = \mathbf{U}_t \mathbf{h}_j + \mathbf{b}_t, \tag{2}$$

where $\mathbf{h}_j \in \mathbb{R}^{d_h}$ is the pretrained node embedding of node j, $\mathbf{v}_j \in \mathbb{R}^{d_x}$ is the transformed feature with type-specific information, and $\mathbf{U}_t \in \mathbb{R}^{d_x \times d_h}$, $\mathbf{b}_t \in \mathbb{R}^{d_x}$. There are different sets of $\{\mathbf{U}_t, \mathbf{b}_t\}$ for input CUIs $\mathcal{E}_x$, target CUIs $\mathcal{E}_y$, and internal CUIs $\mathcal{E}_i$, which are $\{\mathbf{U}_x, \mathbf{b}_x\}$, $\{\mathbf{U}_y, \mathbf{b}_y\}$ and $\{\mathbf{U}_i, \mathbf{b}_i\}$ respectively, and only one of them is used in Eq. (2) according the type of $\mathbf{h}_j$.

3.2.2 Multi-hop Message Passing. In addition to the node type information, another important information that a node embedding should include is the correlation between the node itself and its neighboring entities. To incorporate this information, the first step is to find out the message passing paths that connect input CUI entities $\mathcal{E}_x$ and target CUI entities $\mathcal{E}_y$. For computational efficiency, we define a hyper-parameter K, which denotes the maximum number of hops in a message passing path from $\mathcal{E}_x$ to $\mathcal{E}_y$. The set of valid K-hop paths is defined as:

$$P_k = \{(e_s, r_1, \cdots, r_k, e_d) | (e_s, r_1, e_1), \cdots, (e_{k-1}, r_k, e_d) \in \mathcal{G}\}, (1 \le k \le K) \tag{3}$$

where e_s is the source node, e_d is the destination node, and $r_j (1 \le j \le k)$ is the j-th relation in the path that connects entities e_{j-1} and e_j. Note that e_s and e_d can be any entities in $\mathcal{E}$, which are not limited to $e_s \in \mathcal{E}_x$ and $e_d \in \mathcal{E}_y$. Thus, the set P_k contains all paths from an arbitrary node in input CUIs $\mathcal{E}_x$ to an arbitrary node in target CUIs $\mathcal{E}_y$ with hops equal to k.

Given the path set P_k, which tracks all possible diseases development progress, we now need to use this information to update the embedding of each node e_j in graph $\mathcal{G}$. For different relation r_l, we use a transformation matrix $\mathbf{W}^t_{r_l} \in \mathbb{R}^{d_x \times d_x}$ $(1 \le t \le K)$ to denote how this relation passes the information from source node e_s to e_d. The value of $\mathbf{W}^t_{r_l}$ depends on the distance t from e_s to e_d in any given path $p = (e_s, r_1, \cdots, r_k, e_d) \in P_k$ that passes through e_d. Note that the number of hops k in some path p is smaller than K, we need to introduce padding matrices $W_0^{k+1}, ..., W_0^K \in \mathbb{R}^{d_x \times d_x}$ to maintain the matrix multiplication scale as K for parallel training. Thus, for all paths with k hops in P_k that pass through node e_d, the total updated information they have for e_d is embedded as follows

$$\mathbf{z}^k_d = \sum_{p \in P_k} \text{attn}(p) \cdot W_0^K \cdots W_0^{k+1} W^k_{r_k} \cdots W^1_{r_1} \mathbf{v}_s, (1 \le k \le K), \tag{4}$$

where $\mathbf{v}_s \in \mathbb{R}^{d_s}$ is the embedding of source node e_s in path $p = (e_s, r_1, \cdots, r_k, e_d)$. Note that we introduce a structured relational attention mechanism attn$(\cdot)$, which automatically generates a constant to distinguish the contribution of different path p. The introduction of relational attention mechanism is useful in flexibly selecting important paths of disease development, which is helpful for the prediction interpretation and will be detailed in Section 3.2.3.

In Eq. (4), we define separate transformations for different positions and different relations and also pay attention to the lengths of different paths, which allows us to better learn the correlation between node e_j and others. Now we need to aggregate all $\mathbf{z}^k_j$ with different k to get the final embedding $\mathbf{h}'_j$ for each node e_j in $\mathcal{G}$. Remember that we have learned an EHR representation $\mathbf{s}$ from the EHR encoder, so we use $\mathbf{s}$ to guide the aggregation process with a weighted sum calculated by the bilinear attention mechanism,

$$\mathbf{z}_j = \sum_{k=1}^{K} \text{Softmax}(\text{bilinear}(\mathbf{s}, \mathbf{z}^k_j)) \cdot \mathbf{z}^k_j, \tag{5}$$

where $\mathbf{z}_j \in \mathbb{R}^{d_x}$. Softmax$(\cdot)$ is used to normalize the attention score corresponding to each $\mathbf{z}^k_j$ calculated by function bilinear$(\cdot)$, and function bilinear$(\mathbf{s}, \mathbf{z}^k_j) = \mathbf{s}^\top \mathbf{B} \mathbf{z}^k_j \in \mathbb{R}$ where $\mathbf{B} \in \mathbb{R}^{d_s \times d_x}$ is a learnable matrix.

Now we have learned a feature $\mathbf{z}_j$ containing the information between the node e_j and the rest entities in the personalized graph $\mathcal{G}$. The final step we need to take is using $\mathbf{z}_j$ to update the node embedding $\mathbf{h}_j$,

$$\mathbf{h}'_j = \sigma(\mathbf{T} \cdot \mathbf{h}_j + \mathbf{T}' \cdot \mathbf{z}_j), \tag{6}$$

where $\mathbf{h}'_j \in \mathbb{R}^{d_h}$, $\mathbf{T} \in \mathbb{R}^{d_h \times d_h}$ and $\mathbf{T}' \in \mathbb{R}^{d_h \times d_x}$ are learnable transformation matrix, and σ is an activation function.

3.2.3 Structured Relational Attention. Before moving onto how to use $\mathbf{h}'_j$ for the final prediction, we want to discuss the design of **attn(·)** function in Eq. (4). Recall that function attn(·) is used to calculate the contribution of different paths $p = (e_s, r_1, \cdots, r_k, e_d)$ with the same hop number k. Thus, function attn(·) is useful in finding out important relations between input CUIs $\mathcal{E}_x$ and target CUIs $\mathcal{E}_y$ and interpreting the prediction process.

`MedPath` is flexible in the choice of function attn(·), and we provide two different choices of attn(·) in this paper, i.e. transition matrix-based attention and relational self-attention. These two variants of `MedPath` are referred as `MedPath`-TA and `MedPath`-SA, respectively. The transition matrix-based attention is introduced in [10], which computes the attention weights by a probabilistic graphical model. Although transition matrix-based attention is effective and efficient for distinguishing the importance of different paths, this approach has an assumption that the probability of transition between two relations is fixed. However, in healthcare, different patients usually have different disease development progression, and thus, the transition probability should be dynamically adjusted. To achieve this goal, we propose a new relational self-attention to dynamically model more complex correlation between different relations in different paths, which further helps improve the information flow of multi-hop message passing in `MedPath`.

Transition Matrix-based Attention. Transition matrix-based attention regards the attention score attn(p) of path $p = (e_s, r_1, \cdots, r_k, e_d)$ as the probability of the path p conditioned on $\mathbf{s}$:

$$\text{attn}(p) = \text{probability}(p|\mathbf{s}), \tag{7}$$

which is modeled by a probabilistic graphical model such as conditional random field [15]:

$$\begin{aligned} &\text{probability}(p|\mathbf{s}) \\ &\propto \exp\Big(\mu(\phi(e_s), \mathbf{s}) + \sum_{t=1}^{k} \delta(r_t, \mathbf{s}) + \sum_{t=1}^{k-1} \tau(r_t, r_{t+1}) + \nu(\phi(e_d), \mathbf{s})\Big) \\ &\triangleq \underbrace{\beta(r_1, \cdots, r_k, e_d)}_{\text{Relation Type Attention}} \cdot \underbrace{\gamma(\phi(e_s), \phi(e_d), \mathbf{s})}_{\text{Node Type Attention}}, \end{aligned} \tag{8}$$

where function $\phi(\cdot)$ outputs the node type of the input node. In implementation, functions $\mu(\cdot)$, $\nu(\cdot)$ and $\delta(\cdot)$ are learned by two-layer multilayer perceptrons (MLPs) and $\tau(\cdot)$ by a transition matrix $\in \mathbb{R}^{m \times m}$, where m is the number of relations.

Relational Self-Attention. From Eq. (8), we can see that the probability is defined as the product of relation type attention $\beta(\cdot)$ and node type attention $\gamma(\cdot)$. Function $\beta(\cdot)$ models the importance of a k-hop relation, while function $\gamma(\cdot)$ models the importance of messages from source CUIs to destination CUIs based on $\mathbf{s}$. Motivated by this idea, we propose an improved attn(·) by incorporating self-attention mechanism [27].

For modeling the differences among different patients, instead of using a fixed relation transition matrix $\tau(\cdot)$ as Eq. (8) does, we consider dynamically generating an $m \times k$ score matrix for every relation type at each hop conditioned on $\mathbf{s}$. First, the model takes the EHR representation $\mathbf{s}$ as the input and performs hop-specific transformation for the j-th hop by $\mathbf{a}_j = \mathbf{M}_j\mathbf{s}$, where $\mathbf{M}_j \in \mathbb{R}^{m \times d_s}$ and $\mathbf{a}_j \in \mathbb{R}^m$. We pack all $\mathbf{a}_j$ together and have a matrix $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_k] \in \mathbb{R}^{m \times k}$. We use the self-attention to capture relation-type correlations among different hops within a path,

$$\text{SelfAttention}(\mathbf{A}) = \text{Softmax}(\frac{\mathbf{A}_q\mathbf{A}_k^\top}{\sqrt{d}})\mathbf{A}_v, \tag{9}$$

where matrices $\mathbf{A}_q = \mathbf{M}_q\mathbf{A}$, $\mathbf{A}_k = \mathbf{M}_k\mathbf{A}$, and $\mathbf{A}_v = \mathbf{M}_v\mathbf{A}$ are the query, key and value matrices transformed from $\mathbf{A}$ respectively, where $\mathbf{M}_q, \mathbf{M}_k, \mathbf{M}_v \in \mathbb{R}^{d \times m}$, and d is the number of transformed features.

The output of SelfAttention($\mathbf{A}$) is then mapped back to the original space $\mathbb{R}^m$ by a linear transformation $\mathbf{M}_l \in \mathbb{R}^{m \times d}$ and passed to softmax activation function to generate attention scores for all possible k-hop relations:

$$\beta(r_1, \cdots, r_k, \mathbf{s}) = \text{Softmax}(\mathbf{M}_l \cdot \text{SelfAttention}(\mathbf{A})). \tag{10}$$

For node type attention $\gamma(\phi(e_s), \phi(e_d), \mathbf{s})$, functions $\gamma(\cdot)$ and $\phi(\cdot)$ are both modeled by two-layer MLPs, and node type attention function $\phi(\cdot)$ is different for the source node and the destination node. Given $\beta(r_1, \cdots, r_k, \mathbf{s})$ and $\gamma(\phi(e_s), \phi(e_d), \mathbf{s})$, we can multiply them together and finally have the attention score attn(p) by taking the normalization of the multiplication score.

3.3 Prediction

Following previous sections, we now have the comprehensive EHR representation $\mathbf{s}$ by EHR encoder F_e and the updated personalized knowledge graph with node embeddings $[\mathbf{h}'_1, \cdots, \mathbf{h}'_n]$ obtained by multi-hop message passing. For the final prediction, we firstly apply attentive pooling [23] over all the target CUI entity features to obtain graph embeddings $\mathbf{g}$. Then we concatenate $\mathbf{g}$ and $\mathbf{s}$ to compute the final output by FC($\mathbf{s} \oplus \mathbf{g}$), where FC is the fully connected layer for classification, and $\oplus$ is the concatenation operation. The whole model is trained jointly end-to-end by minimizing the cross-entropy loss between FC($\mathbf{s} \oplus \mathbf{g}$) and the ground truth label.

3.4 Interpretation

One of the benefits of the proposed `MedPath` is that we can decode all k-hop paths ($1 \le k \le K$) and select the path with high attention scores during feed-forward process for model interpretation. From the attentive pooling layer, we can identify the target CUI entity e_y with the highest attention score. Then, using Eq. (5), we can get the optimal k^* with the highest attention scores. Among all k^*-hop paths, the most possible way that the patient will get the target disease is the path p with the maximum attn(p).

4 EXPERIMENTS

In this section, we will first provide the details of our experimental settings and then discuss the results of comparison experiments and ablation studies. Finally, case studies are included to demonstrate

how MedPath interprets the prediction results from the attention weights and personalized knowledge graph.

4.1 Experimental Setup

4.1.1 Datasets. Our experiments are conducted on three EHR datasets collected from real-world claim data. The statistics of these datasets are shown in Table 1. The target disease in these datasets is heart failure, chronic obstructive pulmonary disease (COPD), and kidney disease, respectively. Patients of these diseases normally experience a chronic and progressive condition for a long period.

4.1.2 Baselines. We consider the following baseline models, including LSTM [11], Dipole [17], [6], SAnD [26], RetainEx [14], Timeline [2], LSAN [29], and HiTANet [16], which also can be used as base models of the proposed MedPath-TA and MedPath-SA approaches.

4.1.3 Implementation. MedPath is implemented by PyTorch framework on an NVIDIA Tesla P100 GPU and Intel Xeon E5-2680 CPUs. The parameters are trained by Adam optimizer with the learning rate of 10^{-4} and the mini-batch size is 64. The hidden state numbers in LSTM and GRU are both 256, while they are 128 in the rest methods. As for the graph encoder, the embedding size of each CUI code is 100 and the layer number is 1.

4.1.4 Evaluation Metrics. The evaluation metrics in our task include F1 score and area under the receiver operating characteristic curve (AUC). AUC is the probability that a model ranks a randomly chosen positive case higher than a randomly chosen negative case. As for F1 score, it is the harmonic mean of precision and recall. Precision penalizes false positives while recall penalizes false negatives. Thus, F1 score is a better evaluation for it takes these two aspects into consideration. The higher these scores are, the better performance risk prediction frameworks have.

4.2 Experimental Results

4.2.1 Performance Comparison. The comparison results are shown in Tables 2 and Table 3. For each baseline method, we first test the vanilla model and get the AUC and F1 scores. Then we build on the vanilla model and add a graph neural network layer to obtain the performance of MedPath-TA and MedPath-SA using either $k = 2$ or 3, where k is the length of knowledge path (the influence of different k's on different methods is in the Appendix section). An additional group of comparison using GRAM [5] as base model can be referred to Table 8 in the Appendix section.

Firstly, we can observe the effectiveness of using personalized knowledge by comparing the AUC and F1 scores between the vanilla models and the MedPath variants. Take MedPath-TA as an example. By comparing the results of vanilla models and MedPath-TA from Table 2, we can see that in almost all the settings, MedPath-TA achieves higher AUCs. While from Table 3 in 75% settings, MedPath-TA has higher F1 scores. Notably, the largest improvement in AUC takes place in Dipole, SAnD and RetainEX on the heart failure, COPD and kidney disease datasets, respectively. The largest improvement in F1 scores is taken place in Dipole, RetainEx and RetainEx for heart failure, COPD and kidney disease, respectively. Existing base models emphasize interpreting prediction through attention weights, so MedPath is especially good at coping with existing interpretable frameworks. We also take the average of AUC and F1 score over all base methods for different datasets. The average AUC increment that MedPath-TA has over the vanilla model is 0.033, 0.022 and 0.018 in heart failure, COPD and kidney disease, respectively. The average F1 score increment is 0.045, 0.019 and 0.012 in heart failure, COPD and kidney disease, respectively.

Table 1: Statistics of the used datasets.

Dataset	Heart Failure	COPD	Kidney Disease
Positive Cases	3,080	7,314	7,810
Negative Cases	9,240	21,942	8,430
Average Visits per Patient	30.39	38.74	39.09
Average Codes per Visit	4.24	3.50	4.40
Unique ICD-9 Codes	8,692	10,053	8,802

After determining the effectiveness of graph encoder, the next step is finding out how a better structural relational attention design in graph neural network layer can improve the performance of MedPath. In Section 3.2.3, we have discussed two types of structural relational attentions, i.e. transition matrix-based attention (TA) and relational self-attention (SA), and we distinguish these two MedPath varients by MedPath-TA and MedPath-SA, respectively. Specifically speaking, MedPath-SA can achieve higher AUCs than MedPath-TA in 79.2% settings in Table 2, and it can achieve higher F1 scores than MedPath-TA in two-thirds separate settings in Table 3. Among them, the highest improvement in AUC is achieved in the case of LSTM, HiTANet and SAnD for heart failure, COPD and kidney disease, respectively. The highest F1 score improvement is achieved in the case of LSTM, LSTM and SAnD for heart failure, COPD and kidney disease, respectively. These models cover a broad range of risk prediction models, including RNN-based, attention-based and time aware-based ones. In addition, MedPath-SA achieves higher average AUC and F1 scores on all three datasets. These results show that the use of relational self-attention contributes to better risk prediction performance in various types of risk prediction, and the designed relational self-attention is a better choice than the transition matrix-based attention.

To sum up, the improvement is brought by the personalized knowledge graph feature encoded by the graph neural network module. The new feature embeds the professional medical knowledge in SemMed, which makes the reasoning of the neural networks closer to the reasoning of physicians in real life. With the guidance of the personalized knowledge graph, our model could identify important ICD codes more easily in the records which have plausible relations with the target diseases and help improve the performance.

4.2.2 Significance Test. We also conduct significance test to compare the differences between all the results obtained by vanilla models and MedPath variants. To achieve this goal, we use the Bootstrap method to randomly choose 1,000 testing data and then calculate the values of AUC and F1 score five times. The hypothesis is that the AUC or F1 score means of baseline approaches are the same as those of the proposed MedPath variants. Student's T-test is used with significance level α as 1% to calculate the p-values. From the p-values listed in Tables 2 and 3, we reject the null hypothesis and accept the alternative hypothesis, i.e., true means are totally

Table 2: Performance Comparison (with the p-values of significance test) in terms of AUC. The average AUC scores of our MedPath variants MedPath-TA and MedPath-SA for each dataset are followed by the percentage improvement (↑) over Vanilla models.

Dataset	Heart Failure			COPD			Kidney Disease		
Method	Vanilla	MedPath-TA	MedPath-SA	Vanilla	MedPath-TA	MedPath-SA	Vanilla	MedPath-TA	MedPath-SA
LSTM	0.708	0.716 (1e-10)	**0.739** (6e-10)	0.693	0.703 (4e-9)	**0.707** (7e-9)	0.739	0.762 (4e-10)	**0.774** (1e-10)
Dipole	0.687	0.744 (2e-8)	**0.751** (2e-8)	0.704	0.714 (2e-10)	**0.728** (1e-10)	0.755	0.765 (3e-7)	**0.768** (2e-7)
Retain	0.689	0.733 (2e-8)	**0.735** (5e-8)	0.699	0.723 (6e-10)	**0.730** (6e-10)	0.732	**0.766** (1e-7)	0.764 (3e-7)
SAnD	0.686	0.733 (1e-7)	**0.745** (1e-7)	0.692	0.736 (7e-10)	**0.737** (9e-11)	0.748	0.769 (2e-7)	**0.790** (5e-8)
RetainEx	0.688	0.738 (6e-9)	**0.751** (2e-9)	0.707	**0.746** (2e-9)	0.743 (2e-9)	0.728	0.772 (2e-8)	**0.786** (3e-9)
Timeline	0.705	**0.735** (3e-9)	0.729 (2e-8)	0.698	**0.713** (4e-9)	0.704 (1e-9)	0.756	0.761 (6e-9)	**0.769** (7e-9)
LSAN	0.738	0.729 (9e-8)	**0.745** (1e-7)	0.723	**0.728** (4e-6)	0.720 (2e-6)	0.766	0.765 (9e-7)	**0.782** (5e-8)
HiTANet	0.750	**0.785** (4e-8)	**0.785** (3e-8)	0.752	0.787 (7e-11)	**0.799** (1e-10)	0.792	0.800 (8e-8)	**0.810** (4e-7)
Average	0.706	0.739 (↑4.7%)	**0.748 (↑5.9%)**	0.709	0.731 (↑3.1%)	**0.734 (↑3.5%)**	0.752	0.770 (↑2.4%)	**0.780 (↑3.7%)**

Table 3: Performance Comparison (with the p-values of significance test) in terms of F1 Score. The average F1 scores of our MedPath variants MedPath-TA and MedPath-SA for each dataset are followed by the percentage improvement (↑) over Vanilla models.

Dataset	Heart Failure			COPD			Kidney Disease		
Method	Vanilla	MedPath-TA	MedPath-SA	Vanilla	MedPath-TA	MedPath-SA	Vanilla	MedPath-TA	MedPath-SA
LSTM	0.561	0.582 (2e-10)	**0.611** (1e-9)	0.548	**0.554** (7e-9)	0.550 (1e-8)	0.616	0.646 (3e-9)	**0.661** (3e-9)
Dipole	0.542	0.625 (6e-8)	**0.631** (8e-8)	0.562	0.560 (1e-9)	**0.584** (5e-10)	0.656	**0.657** (1e-6)	0.645 (1e-6)
Retain	0.549	**0.613** (4e-8)	0.612 (1e-7)	0.555	0.570 (2e-9)	**0.575** (1e-9)	0.614	**0.654** (4e-7)	**0.654** (6e-7)
SAnD	0.544	0.605 (2e-7)	**0.609** (3e-7)	0.539	0.568 (1e-9)	**0.590** (3e-10)	0.636	0.633 (9e-7)	**0.670** (3e-7)
RetainEx	0.546	0.625 (4e-9)	**0.640** (2e-9)	0.570	**0.613** (1e-9)	0.612 (1e-9)	0.612	0.665 (4e-8)	**0.682** (2e-8)
Timeline	0.574	**0.614** (5e-9)	0.609 (1e-8)	0.550	**0.570** (6e-9)	0.555 (2e-9)	0.648	0.647 (1e-8)	**0.657** (3e-9)
LSAN	0.623	0.607 (2e-7)	**0.626** (1e-7)	0.574	**0.586** (1e-5)	0.573 (4e-6)	0.661	0.651 (4e-6)	**0.668** (1e-6)
HiTANet	0.647	**0.678** (1e-7)	0.668 (1e-7)	0.637	**0.670** (8e-10)	**0.670** (4e-10)	**0.702**	0.687 (5e-7)	0.688 (8e-7)
Average	0.574	0.619 (↑7.8%)	**0.626 (↑9.1%)**	0.567	0.586 (↑3.4%)	**0.589 (↑3.9%)**	0.643	0.655 (↑1.9%)	**0.666 (↑3.6%)**

different. These results confirm that MedPath is significantly better than vanilla baselines.

4.2.3 Model Interpretability. In addition to performance improvement, the other reason that we incorporate medical paths is to provide explicit explanations to interpret the prediction results and highlight the influential relationship between different symptoms that will eventually lead to the target disease. In this part, we verify our claim by giving a case study on how to interpret the risk prediction with MedPath as shown in Table 4. The target disease in this case study is heart failure, and the first row shows the EHR data of the patient, which consist of 6 visits in total. Each visit is described with several diagnosis codes. For the ease of illustration, we only show the 2-hop paths in the personalized knowledge graph. In Table 4, we select three 2-hop paths with the highest attention weights and one 2-hop path with the lowest attention weight learned by MedPath-SA with LSTM. Here we can see how MedPath explicitly illustrates the correlation between the EHRs and the target disease. Take ICD-9 code "244.9" (hypothyroidism) as an example, which has the largest attention weight. As the second row shows, the relation between hypothyroidism and heart failure disease is that it causes hypertensive disease and that can further lead to left heart failure. If we want to have supportive medical findings on how hypothyroidism can lead to hypertensive disease and how hypertensive disease can finally lead to left heart failure, we can see Evidence E1 and Evidence E2 in Table 4. They are the abstract sentences extracted by SemMed from reliable medical publications. Let us move to the third path, which is from fatigue to left heart failure and is not highly related to the target disease. Thus, the weight of this path is much lower than those of the top two paths. As for the path with the lowest attention weight, we can see that the source entity and the destination entity are all heart failure, which does not provide any useful information on the relation between the EHR example and target disease. Thus, it is correct for MedPath to give it a zero weight. This case study can clearly show the reasonableness of the proposed MedPath with reliable explanations on prediction results. Besides, two more case studies are shown in Tables 6 and 7 in Appendix Section A.

4.2.4 Ablation Study. Now we investigate our framework design by the ablation study. We break down our MedPath layer into type-specific transformation (Eq. (2)), node-type attention (i.e., $\gamma(\cdot)$ function in Eq. (8)), and relational self-attention components to see how

Table 4: Case study results of heart failure for showing the explicit interpretability that MedPath has.

EHR Data	**Visit 1**: 250.02 (Diabetes mellitus); **Visit 2**: 585.9 (Chronic kidney disease) and 780.79 (Fatigue); **Visit 3**: 244.9 (Hypothyroidism), 272.4 (Hyperlipidemia), and 401.1 (Benign essential hypertension); **Visit 4**: 585.9 (Chronic kidney disease); **Visit 5**: 585.9 (Chronic kidney disease); **Visit 6**: 585.9 (Chronic kidney disease) and 244.9 (Hypothyroidism)	
1st Highest Attention Weighted Path	Weight: 0.0189	Hypothyroidism $\xrightarrow[E1]{CAUSES}$ Hypertensive disease $\xrightarrow[E2]{CAUSES}$ Left heart failure
	Evidence E1	*Animal studies suggest that hypertension leads to cardiac tissue hypothyroidism a condition that can by itself lead to heart failure.*
	Evidence E2	*Left ventricular failure in some SA/OHS patients may be the result of hypertensive cardiac disease.*
2nd Highest Attention Weighted Path	Weight: 0.0178	Hyperlipidemia $\xrightarrow[E3]{CAUSES}$ Hypertensive disease $\xrightarrow[E4]{CAUSES}$ Left heart failure
	Evidence E3	*A literature search indicates that Anglo-Saxon countries report alarming hyperplastic changes particularly in the liver blood clots hyperlipidemia leading to high blood pressure porphyria atypical leiomyomas and cervical hyperplasia.*
	Evidence E4	*Left ventricular failure in some SA/OHS patients may be the result of hypertensive cardiac disease.*
3rd Highest Attention Weighted Path	Weight: 0.0150	Fatigue $\xrightarrow[E5]{CAUSES}$ Cessation of life $\xrightarrow[E6]{CAUSES}$ Left heart failure
	Evidence E5	*In light of the magnitude of this sleep debt it is not surprising that fatigue is a factor in 57% of accidents leading to the death of a truck driver and in 10% of fatal car accidents and results in costs of up to 56 billion dollars per year.*
	Evidence E6	*Though rare death due to myocardial stunning and LV power failure can occur during ICD insertion.*
One of the Lowest Attention Weighted Path	Weight: 0.0000	Heart failure $\xrightarrow[E7]{CAUSES}$ Hypertensive disease $\xrightarrow[E8]{CAUSES}$ Left heart failure
	Evidence E7	*These findings suggest that the ATF3 activator tBHQ may have therapeutic potential for the treatment of pressure-overload heart failure induced by chronic hypertension or other pressure overload mechanisms.*
	Evidence E8	*Left ventricular failure in some SA/OHS patients may be the result of hypertensive cardiac disease.*

Table 5: Ablation study results of removing each component in MedPath-SA using LSTM as the encoder.

Dataset	Heart Failure	COPD	Kidney
Methods	AUC	AUC	AUC
MedPath-SA (k = 3)	**0.739**	**0.707**	**0.774**
- type specific transformation	0.707	0.701	0.759
- node type attention	0.714	0.704	0.746
- relational self attention	0.724	0.701	0.755

they influence the model performance. Without loss of generalization, we use LSTM as the base model of MedPath-SA for analysis in this part. After removing each component individually, from Table 5 we can see that the model performance in AUC decrease for at least 1.5%, 0.3% and 1.5% in heart failure, COPD and kidney disease, respectively. These results conclude our ablation study and show that each component is essential for the design of MedPath.

5 RELATED WORK

5.1 Health Risk Prediction

Health risk prediction task is a developing area that attracts significant attention due to its great potential in real-world medical application. A line of risk prediction works mainly focuses on how to better utilize deep learning techniques, which adopt RNNs like gated recurrent units (GRUs) [7], long-short-term memory (LSTM) [11], and self attention-based networks [27]. The representative works include Retain [6], Dipole [17], SAnD [26], and LSAN [29]. The development in this direction is driven by the intuition that representations learned by better deep learning networks are helpful for prediction results. In addition, since EHR data provide the time information of each visit, risk prediction works also use deep learning frameworks to incorporate time information to improve the representation learning. Representative works of using time information include Time-aware LSTM [2], RetainEx [14], Timeline [1], and HiTANet [16], which believe that the time information on visits can indicate the individual disease progression of different patients.

Another line of works [5, 20, 30, 31] leverage medical ontology or general knowledge graphs to improve the feature learning for risk prediction. However, we should note that the use of knowledge graph in existing knowledge-graph-based models like GRAM [5] and KAME [20] is for improving the medical code embedding used for all patients. Unlike them, MedPath uses a personalized knowledge graph extracted from SemMed to learn the disease progression information for each individual patient. This way of using knowledge graph is helpful for incorporating personalized information for more accurate risk prediction.

5.2 Graph Neural Network

Graph neural network (GNN) is a special kind of neural network used for graph data, which learns the features of nodes by aggregating information passed from neighboring ones. The GNN technique

is first fueled by the work of Graph Convolutional Network (GCN) by Kipf et al. [13], which builds up the foundation of how to expand the convolution operation on graph data. After that, new types of GNN like Graph Attention Network (GAT) [28] and R-GCN [24] are proposed to propagate neighboring information with graph attention and efficiently handle multi-relational data, respectively.

However, the message-passing process in GCN and R-GCN is single-hop, which cannot handle graphs like the knowledge graph in SemMed to capture more complex relations between nodes. Our work is motivated by a more up-to-date GNN called multi-hop graph relation network (MHGRN) [10] that can provide the multi-hop inference between symptom nodes and target disease nodes. Specifically, the MedPath variant called MedPath-TA include MHGRN as the graph encoder to learn the embeddings of personalized graphs through multi-hop messages. In addition, we develop on the idea of MedPath-TA and provide another variant called MedPath-SA to handle multi-relation learning in different hops.

6 CONCLUSIONS

In this paper, we introduce a new general framework, called MedPath, to enable existing health risk prediction methods to incorporate personalized information and provide explicit interpretation for predictions. To find out the correlations between symptoms and target diseases, MedPath first extracts a personalized knowledge graph (PKG) for each patient from the SemMed web which contains a giant medical knowledge graph. After that, MedPath finds out all possible disease progression paths from the PKG and uses them to learn a personalized embedding to augment the base risk prediction models for the final prediction. Since the graph neural network encoder for PKG assigns attention weights on disease progression paths instead of independent medical codes, MedPath is able to provide reliable explicit explanations in the testing phrase by showing paths in PKG with high attention weights. Experimental results show that our model is able to improve the performance of existing models in terms of both F1 score and AUC. More importantly, in case studies, we confirm that MedPath can provide explicit explanations for its prediction through paths in PKG. In the future, we would like to work on how to allow risk prediction models to infer new emerging links between conditions.

REFERENCES

[1] Tian Bai, Shanshan Zhang, Brian L Egleston, and Slobodan Vucetic. 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.

[2] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. 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.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*. 2787–2795.

[4] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 432–440.

[5] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 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*. ACM, 787–795.

[6] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*. 3504–3512.

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[8] Kevin Donnelly. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics* 121 (2006), 279.

[9] Patrick Ernst, Amy Siu, and Gerhard Weikum. 2015. Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC bioinformatics* 16, 1 (2015), 157.

[10] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering. *arXiv preprint arXiv:2005.00646* (2020).

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[12] Halil Kilicoglu, Marcelo Fiszman, Alejandro Rodriguez, Dongwook Shin, A Ripple, and Thomas C Rindflesch. 2008. Semantic MEDLINE: a web application for managing the results of PubMed Searches. In *Proceedings of the third international symposium for semantic mining in biomedicine*, Vol. 2008. 69–76.

[13] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[14] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. 2018. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 299–309.

[15] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289. http://dl.acm.org/citation.cfm?id=645530.655813

[16] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 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.

[17] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 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.

[18] Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. 2018. Risk prediction on electronic health records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1910–1919.

[19] Fenglong Ma, Yaqing Wang, Houping Xiao, Ye Yuan, Radha Chitta, Jing Zhou, and Jing Gao. 2018. A general framework for diagnosis prediction via incorporating medical code descriptions. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1070–1075.

[20] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 743–752.

[21] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 30–41.

[22] Thomas C Rindflesch, Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat, and Dongwook Shin. 2011. Semantic MEDLINE: An advanced information management application for biomedicine. *Information Services & Use* 31, 1-2 (2011), 15–21.

[23] Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. *arXiv preprint arXiv:1602.03609* (2016).

[24] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*. Springer, 593–607.

[25] Robert Sedgewick. 2001. *Algorithms in c, part 5: graph algorithms*. Pearson Education.

[26] Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. 2017. Attend and diagnose: Clinical time series analysis using attention models. *arXiv preprint arXiv:1711.03905* (2017).

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[28] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[29] Muchao Ye, Junyu Luo, Cao Xiao, and Fenglong Ma. 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 and Knowledge Management*.

[30] Changchang Yin, Rongjian Zhao, Buyue Qian, Xin Lv, and Ping Zhang. 2019. Domain Knowledge guided deep learning with electronic health records. In *2019*

IEEE International Conference on Data Mining (ICDM). IEEE, 738–747.

[31] Xianli Zhang, Buyue Qian, Yang Li, Changchang Yin, Xudong Wang, and Qinghua Zheng. 2019. KnowRisk: An Interpretable Knowledge-Guided Model for Disease Risk Prediction. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1492–1497.

Appendix

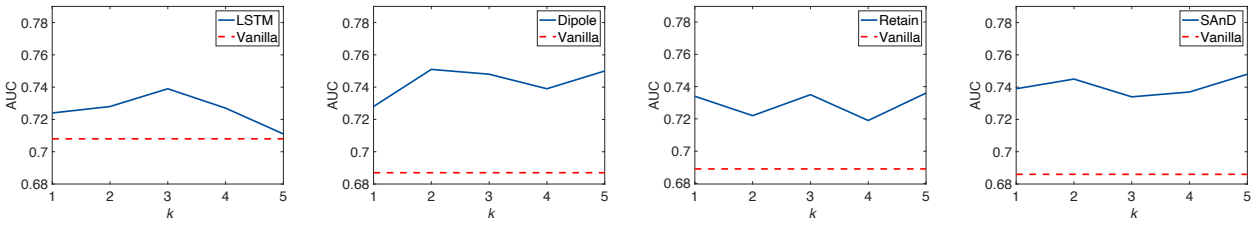
A. Another Two Case Studies. Here we provide two more examples to verify the explicit interpretability that MedPath has. For the ease of analysis, we only extract the 2-hop medical progression paths. We first analyze the positive COPD case as shown in Table 7. In this case, the patient has 6 visits, and each visit is described by at least 3 ICD-9 codes. As shown in the second part of Table 7, MedPath finds out three disease progression paths that are most likely lead to the COPD disease. The path having the highest attention weights starts from the symptom of gastroesophageal reflux disease ("530.81") in the first visit, which will affect the aspiration action of the patient and finally lead to the bronchitis disease. Another two possible paths also end with the bronchitis disease, a common condition of the COPD disease. One of them starts from the coronary arteriosclerosis ("414.00" in Visit 2, 5 and 6) that reduces blood flow in the heart and affects inflammation, while the other starts from the symptom of fatigue ("780.79" in Visit 5) which affects the Ammonia inside the body. The case study example of kidney disease is shown in Table 7. The patient in the case study has 8 visit records. The first thing we should note is that the symptom of hyperlipidemia ("272.4") is recorded in 5 visits, and the path starting from hyperlipidemia is assigned with the highest weight by MedPath. The path with the second highest attention weight starts from rotator cuff syndrome ("726.1" in Visit 1) which causes the atrophic symptom and eventually the kidney failure, and we can have similar analysis on the path with the third highest attention weights. From the weight assignment results, we can see that MedPath can assign high weights to symptoms that occur repeatedly if they are correlated to the target disease, as evidenced by the hyperlipidemia symptom.

B. Discussion for k Selection. As shown in Figure 2, Figure 3 and Figure 4, for each base model with relational self-attention, we increase k from 1 to 5 and compare the AUC scores in all test datasets. These figures shows that among 24 different testing settings, we can obtain the best AUCs when we choose k to be either 2 or 3 in 16 (two-thirds) cases. In addition to better AUC performance, selecting a relatively smaller k also provide straightforward interpretation for risk prediction. Therefore, we recommend tuning k to be 2 or 3.

C. Performance Comparison with GRAM as the base model. In additional to the base models in Table 2, we also use GRAM as the baseline. From Table 8, we can see that in terms of AUC and F1 score, using MedPath framework can help us achieve improvement in 2 out 3 datasets. These results show that MedPath is model-agnostic and can also bring improvements to models using knowledge graph.

Table 6: Case study results of COPD for showing the explicit interpretability that MedPath has.

EHR Data	**Visit 1**: 719.41 (Shoulder joint pain), 530.81 (Gastroesophageal reflux disease) and 272.4 (Hyperlipidemia); **Visit 2**: 708.0 (Allergic urticaria), 272.4 (Hyperlipidemia) and 414.00 (Coronary Arteriosclerosis); **Visit 3**: 413.9 (Angina Pectoris), 272.4 (Hyperlipidemia), 786.50 (Chest Pain) and 425.4 (Cardiomyopathies); **Visit 4**: 426.3 (Left bundle branch block), 401.9 (Essential Hypertension) and 413.9 (Angina Pectoris); **Visit 5**: 786.50 (Chest Pain), 414.00 (Coronary Arteriosclerosis) and 780.79 (Fatigue); **Visit 6**: 300.00 (Anxiety state), 414.00 (Coronary Arteriosclerosis) and 272.4 (Hyperlipidemia)	
1st Highest Attention Weighted Path	Weight: 0.0153	Gastroesophageal reflux disease $\xrightarrow[E1]{AFFECTS}$ Aspiration-action $\xrightarrow[E2]{CAUSES}$ Bronchitis
	Evidence E1	*Gastroesophageal reflux, gastroparesis and achalasia are all associated with aspiration.*
	Evidence E2	*The absence of LLM in 29 control infants suggest that the aspiration may be one cause of recurrent bronchitis in infants.*
2nd Highest Attention Weighted Path	Weight: 0.0118	Coronary Arteriosclerosis $\xrightarrow[E3]{AFFECTS}$ Inflammation $\xrightarrow[E4]{CAUSES}$ Bronchitis
	Evidence E3	*Epicardial fat (EF) is a visceral fat deposit, located between the heart and the pericardium, which shares many of the pathophysiological properties of other visceral fat deposits, It also potentially causes local inflammation and likely has direct effects on coronary atherosclerosis.*
	Evidence E4	*It is speculated that an initial respiratory insult such as viral infection disrupts normal surface morphology and ciliary function, which leads to chronic self-perpetuating inflammation with the formation of bacterial biofilms, leading to PBB.*
3rd Highest Attention Weighted Path	Weight: 0.0118	Fatigue $\xrightarrow[E5]{AFFECTS}$ Ammonia $\xrightarrow[E6]{CAUSES}$ Bronchitis
	Evidence E5	*Serum levels of urea nitrogen (SUN), triglyceride fatty acids (TG), lactate dehydrogenase (LDH), lactic acid (LA), ammonia and hepatic glycogen (HG) were also examined for potential mechanisms underlying the anti-fatigue effect of RPL extracts.*
	Evidence E6	*Acute lung injuries caused due to inhalation of toxic irritant gases such as ammonia, chlorine, hot smoke and burning plastic fumes predominantly affect the airways, causing tracheitis, bronchitis, and other inflammatory responses.*
One of the Lowest Attention Weighted Path	Weight: 0.0000	Agonists $\xrightarrow[E7]{PART\_OF}$ Patients $\xrightarrow[E8]{AFFECTS}$ Bronchitis
	Evidence E7	*Treatment strategy for elderly diabetic patient with insulin or GLP-1 receptor agonist.*
	Evidence E8	*However, in half of the patients (15 cases) the cause was obscure although it was associated with sinusitis, bronchitis or bronchiectasis (Young\\'s syndrome).*



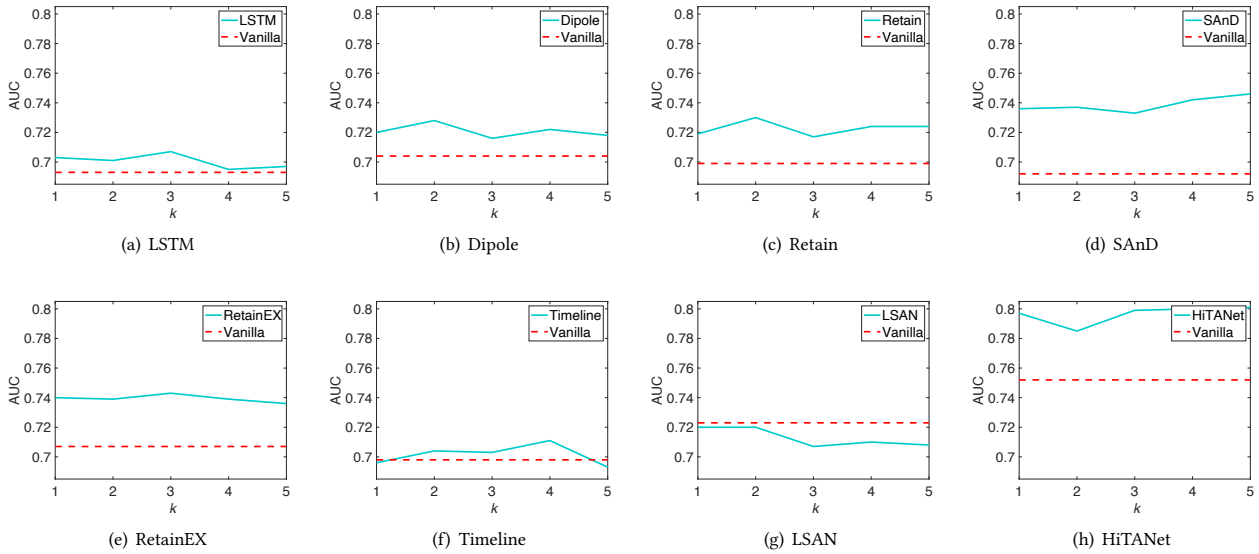
(a) LSTM (b) Dipole (c) Retain (d) SAnD

(e) RetainEX (f) Timeline (g) LSAN (h) HiTANet

Figure 2: Comparison of AUC values with different k in different MedPath-SA methods on the heart failure dataset.

Table 7: Case study results of kidney disease for showing the explicit interpretability that MedPath has.

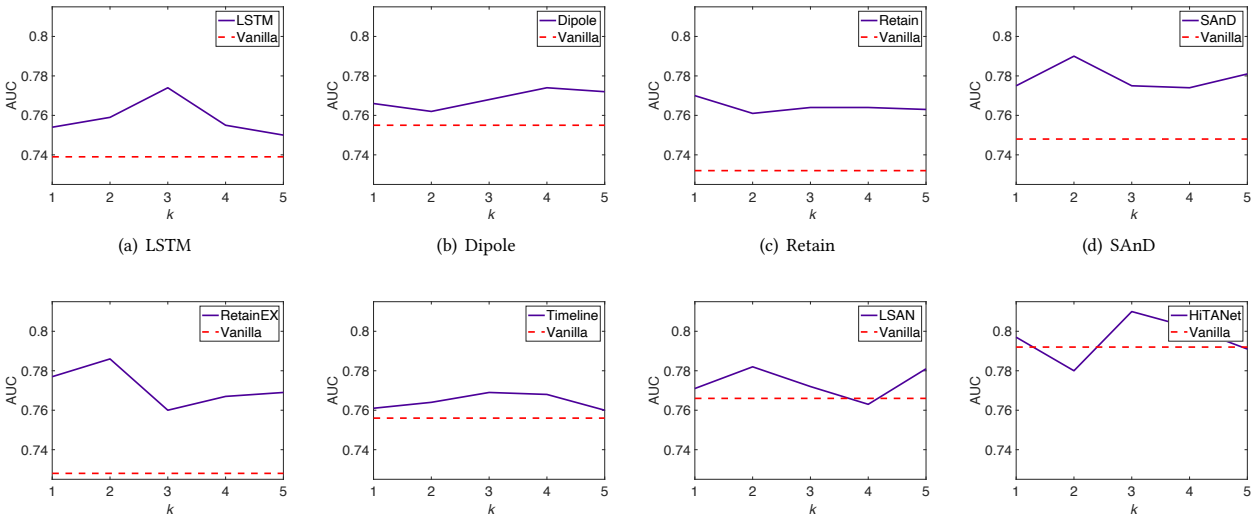
EHR Data	**Visit 1**: 272.4 (Hyperlipidemia), 726.1 (Rotator cuff syndrome) and 401.1 (Benign essential hypertension); **Visit 2**: 401.1 (Benign essential hypertension), 272.4 (Hyperlipidemia) and 794.31 (EEG abnormal); **Visit 3**: 715.00 (Generalized osteoarthritis), 272.4 (Hyperlipidemia); **Visit 4**: 401.1 (Benign essential hypertension), 272.4 (Hyperlipidemia); **Visit 5**: 784.0 (Headache); **Visit 6**: 401.1 (Benign essential hypertension), 272.4 (Hyperlipidemia); **Visit 7**: 719.7 (Difficulty walking), 734 (Flat foot); **Visit 8**: 734 (Flat foot)	
1st Highest Attention Weighted Path	Weight: 0.0152	Hyperlipidemia $\xrightarrow[E1]{CAUSES}$ Hypertensive disease $\xrightarrow[E2]{AFFECTS}$ Kidney Failure
	Evidence E1	*A literature search indicates that Anglo-Saxon countries report alarming hyperplastic changes, particularly in the liver, blood clots, hyperlipidemia leading to high blood pressure, porphyria, atypical leiomyomas and cervical hyperplasia.*
	Evidence E2	*Various factors may play a role in the pathogenesis of hypertension in chronic renal failure.*
2nd Highest Attention Weighted Path	Weight: 0.0133	Rotator cuff syndrome $\xrightarrow[E3]{CAUSES}$ Atrophic $\xrightarrow[E4]{CAUSES}$ Kidney Failure
	Evidence E3	*The RCT made by transecting the supraspinatus (SSP) tendon resulted in atrophy of the SSP muscle.*
	Evidence E4	*Tubular atrophy in the pathogenesis of chronic kidney disease progression.*
3rd Highest Attention Weighted Path	Weight: 0.0119	Headache $\xrightarrow[E5]{CAUSES}$ Magnetic Resonance Imaging $\xrightarrow[E6]{DIAGNOSES}$ Kidney failure
	Evidence E5	*Visual failure poor growth or headache led to MRI diagnosis of CP.*
	Evidence E6	*These findings indicate that DTI MRI may be able to evaluate RF in CKD by DN.*
One of the Lowest Attention Weighted Path	Weight: 0.0000	Renal disease $\xrightarrow[E7]{AFFECTS}$ Homeostasis $\xrightarrow[E8]{AFFECTS}$ Kidney Failure
	Evidence E7	*Renal sympathetic nerve activity has an important role in renal disease-associated hypertension and in the modulation of fluid homeostasis.*
	Evidence E8	*Calcium phosphorus and magnesium homeostasis is altered in chronic kidney disease (CKD).*



(a) LSTM (b) Dipole (c) Retain (d) SAnD

(e) RetainEX (f) Timeline (g) LSAN (h) HiTANet

Figure 3: Comparison of AUC values with different k in different MedPath-SA methods on the COPD dataset.



(a) LSTM (b) Dipole (c) Retain (d) SAnD

(e) RetainEX (f) Timeline (g) LSAN (h) HiTANet

Figure 4: Comparison of AUC values with different k in different MedPath-SA methods on the kidney disease dataset.

Table 8: Performance Comparison with GRAM as the base model.

Dataset	Heart Failure			COPD			Kidney Disease		
Metric	Vanilla	MedPath-TA	MedPath-SA	Vanilla	MedPath-TA	MedPath-SA	Vanilla	MedPath-TA	MedPath-SA
AUC	0.748	**0.751**	0.749	0.722	0.741	**0.744**	**0.780**	0.775	0.778
F1 Score	0.628	**0.639**	0.632	0.582	**0.602**	0.600	**0.677**	0.659	0.666