CAMP: Co-Attention Memory Networks for Diagnosis Prediction in Healthcare

Jingyue Gao[1,3], Xiting Wang[2], Yasha Wang[1,4,*], Zhao Yang[1,3], Junyi Gao[1], Jiangtao Wang[5], Wen Tang[6], Xing Xie[2]
[1]Key Laboratory of High Confidence Software Technologies, Ministry of Education
[2]Microsoft Research Asia; [3]School of Electronics Engineering and Computer Science, Peking University
[4]National Engineering Research Center of Software Engineering, Peking University
[5]School of Computing and Communications, Lancaster University; [6]Peking University Third Hospital
[*]The corresponding author, email: wangyasha@pku.edu.cn

***Abstract*—Diagnosis prediction, which aims to predict future health information of patients from historical electronic health records (EHRs), is a core research task in personalized healthcare. Although some RNN-based methods have been proposed to model sequential EHR data, these methods have three major issues. First, they cannot capture fine-grained progression patterns of patient health conditions. Second, they do not consider the mutual effect between important context (e.g., patient demographics) and historical diagnosis. Third, the hidden state vectors in RNNs are hard to interpret, which leads to trust issues. To tackle these challenges, we propose a model called Co-Attention Memory networks for diagnosis Prediction (CAMP), which tightly integrates historical records, fine-grained patient conditions, and demographics with a three-way interaction architecture built on co-attention. Our model augments RNNs with a memory network to enrich the representation capacity. The memory network enables analysis of fine-grained patient conditions by explicitly incorporating a taxonomy of diseases into an array of memory slots. We design the memory slots to ensure interpretability and instantiate the READ/WRITE operations of the memory network so that the memory cooperates effectively with the patient demographics through co-attention. Experiments and a case study on real-world datasets demonstrate that CAMP consistently performs better than state-of-the-art methods in terms of prediction accuracy and is highly interpretable.**

***Index Terms*—diagnosis prediction, memory networks, attention mechanism, healthcare informatics**

I. INTRODUCTION

Nowadays, Electronic Health Record (EHR) systems are widely adopted to record longitudinal patient health data such as diagnosis, medications, and procedures. The availability of massive amounts of EHR data enables the possibility of clinical predictive tasks [1], [2]. Predicting future diagnosis based on patient's historical records of diagnosis, i.e., *diagnosis prediction* [3], [4], has become a cornerstone of personalized healthcare. This task attracts considerable attention in both industry and the research community because of their importance in need anticipation and precision medicine [4], [5]. Although there is broad consensus on its importance, diagnosis prediction is challenging due to the sequential, high-dimensional, and noisy nature of EHR data.

With recent advances in deep learning, many studies on diagnosis prediction adopt Recurrent Neural Networks (RNNs) to model sequential EHR data. For example, Choi et al. [3] apply RNNs on reversed diagnosis sequences and Ma et al. [6] use Bidirectional RNNs for further improvement. Recently, researchers have incorporated taxonomies of diseases into RNNs [7], [8]. These methods have achieved encouraging prediction accuracy due to their ability to capture dynamic patient conditions and estimate the likelihood of future diagnosis. However, they cannot effectively address the following three challenges in diagnosis prediction.

C1: It is difficult to capture fine-grained progression patterns of patient conditions. The health conditions of a patient can be complicated: diseases are correlated with each other and there may be long-term dependencies between diseases of different categories [6]. To effectively model complex patient health conditions, we need to perform fine-grained analysis on the relationships between the diseases and their attributes (e.g., categories). However, RNNs tend to focus more on short-term memories [9], [10] and would forcefully compress historical records into one hidden state vector. Such highly abstractive features constrain the representation power of RNNs and make it difficult for RNNs to preserve fine-grained information of diagnosed diseases and long-term patient health conditions.

C2: Existing methods cannot model the mutual effect between important context and historical records. Patient demographics are considered important context in the domain of diagnosis prediction [4], [11]. However, how to model the mutual effect between patient demographics and their diagnosed diseases has not been explored, which limits the accuracy of existing methods.

C3: The third major challenge pertains to model interpretability. For medical diagnosis, predictions will not be acted upon based on only blind faith, as the consequences can be devastating [12]. Thus, it is vital that we build an interpretable model so that medical experts can diagnose the model and decide if it is trustworthy. Most existing methods fail to meet this criterion, as the hidden vector representation in RNNs is difficult to interpret [4], [6]. Recently, there have been some efforts in explaining which visits or disease categories are more important in prediction [4], [6], [8]. However, such interpretations are insufficient for understanding the working

mechanism of the prediction model. It is desirable that the fined-grained information used for prediction can be analyzed, connected with the important context (e.g., patient demographics), and verified by medical experts.

Based on these observations, we propose a model called **C**o-**A**ttention **M**emory networks for diagnosis **P**rediction (**CAMP**[1]) , which enhances the **prediction accuracy** and **interpretability** of diagnosis prediction by addressing these three challenges.

The framework of our model is shown in Figure 1. We design a three-way interaction neural architecture built upon co-attention to tightly integrate historical records, fine-grained patient conditions, and demographics. We enable the analysis of fine-grained patient conditions by explicitly incorporating taxonomies of diseases into the framework and memorizing the knowledge contained in the taxonomies with Key-Value Memory Networks (KV-MNs) [13]. Instead of relying on a compressed vector, KV-MNs store different categories of information separately in different memory slots, which enriches the representation capacity compared with RNNs [14], [15]. We elaborately design the memory slots and the READ/WRITE operations of the memory network so that the KV-MN can 1) effectively model the disease categories and their relationships (e.g., connections through ancestors) to capture fine-grained dynamic health conditions of patients (**C1**); 2) cooperate with patient demographics in a mutual enhancement way through a co-attention mechanism (**C2**); and 3) provide meaningful fine-grained interpretations on patient health conditions (**C3**).

To demonstrate the effectiveness of CAMP, we conduct two numerical experiments and one case study on real-world EHR datasets. The experiment on prediction accuracy demonstrates that CAMP consistently outperforms state-of-the-art methods on two datasets in terms of different evaluation criteria. Detailed analysis of CAMP validates the effectiveness of different components and shows that our method is more accurate than competitive baselines with different hyper-parameter settings. The case study illustrates how CAMP can be used to provide meaningful interpretations on fine-grained progression patterns of patient conditions and how our model relates the patient conditions with patient demographics. The findings in the case study have been positively confirmed by a clinical expert.

The rest of this paper is organized as follows. In Section II, we introduce the problem definition of diagnosis prediction. The model of CAMP is proposed and detailed in Section III. Section IV shows experimental results on two real-world EHR datasets. Studies related to our work are reviewed in Section V and Section VI summarizes this work.

II. Problem Definition

We define the problem of diagnosis prediction as follows. For simplicity, all algorithms will be presented for one patient.

Input. For each patient, the input data of our model consists of a sequence of his/her historical records $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t)$, patient demographics $\mathbf{p}$ and a given taxonomy of diseases $\mathcal{G}$.

- The **historical records** of a patient contain multiple visits. Each visit $\mathbf{x}_j \in \{0,1\}^{|\mathcal{C}|}$, $j \in [1,t]$ is a multi-hot binary vector. Here $|\mathcal{C}|$ denotes the number of diseases and $\mathcal{C} = \{c_1, c_2, ..., c_{|\mathcal{C}|}\}$ is the set of entire disease codes from the EHR. $x_{j,i} = 1$ indicates that the patient was diagnosed with disease c_i in the j-th visit.
- The **patient demographics** $\mathbf{p}$ consists of patient characteristics such as age and gender, which are often recorded in the EHR. Let $\mathbf{p} \in \{0,1\}^r$ denote a multi-hot vector indicating demographics of the patient. Following [4], we construct $\mathbf{p}$ by discretizing each attribute (e.g., divide age into several age groups), representing the discretized attributes by using one-hot vectors, and concatenating these vectors.
- The **disease taxonomy** $\mathcal{G}$ expresses the hierarchy of disease concepts in the form of a parent-child relationship, where diseases in $\mathcal{C}$ form the leaf nodes. As shown in Figure 1, a parent node (e.g., *viral infection*) in $\mathcal{G}$ is a disease category that summarizes the diseases described by its children (e.g., *HIV infection* and *Hepatitis*). All nodes in $\mathcal{G}$ form the set $\mathcal{D} = \mathcal{C} + \mathcal{C}'$, where $\mathcal{C}' = \{c_{|\mathcal{C}|+1}, ..., c_{|\mathcal{C}|+|\mathcal{C}'|}\}$ consists of all ancestor nodes. The L nodes at the highest hierarchical level of $\mathcal{G}$ represent the most general categories of diseases (e.g., *Infectious and parasitic diseases*). We call these nodes top-level categories. We build $\mathcal{G}$ by using well-organized taxonomies of diseases (e.g., ICD[2] and CCS[3]).

Output: Given the historical records of a patient, his/her demographics, and a disease taxonomy, the output of our model is the predicted diagnosis of the next visit: $\tilde{\mathbf{x}}_{t+1}$.

III. The Proposed Model

In this section, we first introduce the model overview. Then, we describe the design of the two major components in CAMP and illustrate how the components can be jointly optimized.

A. Model Overview

Figure 1 shows the framework of CAMP, which is a three-way interaction architecture that tightly integrates historical records, fine-grained patient conditions, and demographics. In particular, CAMP predicts future diagnosis with two major components.

Memory-augmented sequential encoder. This component captures fine-grained dynamic health conditions of a patient by augmenting RNN-based models with external memory networks. In our design, the RNN models short-term sequential patient conditions. The memory network encodes fine-grained long-term patient conditions by incorporating knowledge from the disease taxonomy. The two parts cooperate through the memory attention and READ/WRITE operations of the memory network.

Co-attention-based mutual enhancement. Considering that patient health conditions and resistance to potential diseases are closely associated with demographics, CAMP

[1]The source code is available at CAMP.

[2]https://www.cdc.gov/nchs/icd/index.htm

[3]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

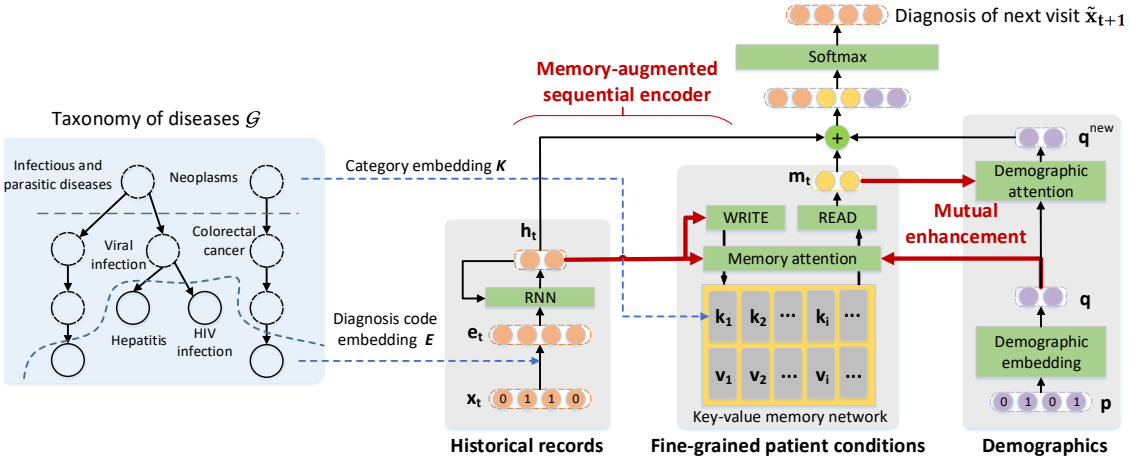

Figure 1: Framework of the proposed model for diagnosis prediction.

employs a co-attention mechanism to mutually improve the representations of patient health conditions and demographics. READ and WRITE operations on the memory matrix are performed attentively and we specifically consider the representation of demographics $\mathbf{q}$ in determining the attention weights over different memory slots. Moreover, we further enhance $\mathbf{q}$ to derive $\mathbf{q}^{new}$ via an attention layer conditioned on the memory network. In this way, the demographic embeddings are improved according to specific memory-based context.

With above components, we obtain $\mathbf{h}_t$, which captures patient conditions in the short term, $\mathbf{m}_t$ that captures fine-grained information over the long term, and the enhanced demographics representation $\mathbf{q}^{new}$. The three representations are jointly considered in the final prediction.

B. Memory-Augmented Sequential Encoder

Our memory-augmented sequential encoder consists of three parts: diagnosis embedding, the RNN, and the memory network.

1) Diagnosis Embedding: The goal of diagnosis embedding is to encode hierarchical medical knowledge in the representations of diseases and their categories. The enriched embeddings help to handle data insufficiency and enhance model accuracy. Given taxonomy $\mathcal{G}$ (shown in Figure 1), we learn robust embeddings for each node in $\mathcal{G}$ by using a state-of-the-art method, GRAM [7]. GRAM represents a node as a combination of itself and its ancestors in $\mathcal{G}$ via a graph-based attention mechanism. With embeddings of all nodes available, we construct two embedding matrices:

- Diagnosis code embedding matrix $\mathbf{E} \in \mathbb{R}^{d_1 \times |\mathcal{C}|}$, which contains embeddings of all leaf nodes (i.e., diseases).
- Category embedding matrix $\mathbf{K} \in \mathbb{R}^{d_1 \times L}$, which contains embeddings of all top-level disease categories.

Here, d_1 is the embedding size. L is the number of top-level nodes.

2) RNN: Recurrent neural networks (RNNs) have been proven effective in modeling the temporal dependency in sequences. To tackle the problem of vanishing gradient, Long-Short Term Memory (LSTM) [16] and Gated Recurrent Unit (GRU) [17] have been proposed as improved variants. We choose GRU here since it can achieve similar performance as LSTM with fewer parameters.

Let d_2 denote the hidden size of GRU, current hidden state vector $\mathbf{h}_t \in \mathbb{R}^{d_2}$ of GRU can be computed recursively:

$$\mathbf{h}_t = \text{GRU}(\mathbf{h}_{t-1}, \mathbf{e}_t; \Theta), \tag{1}$$

where $\text{GRU}(\cdot)$ is the GRU unit, $\mathbf{e}_t = \mathbf{E}\mathbf{x}_t$ is the embedded diagnosis of t-th visit, $\mathbf{h}_{t-1}$ denotes the previous hidden state vector, and Θ represents all parameters of the GRU unit. Researchers have shown that RNNs tend to capture disease progression in the short term and fail to remember patient health conditions over the long term [18]. Thus, we consider $\mathbf{h}_t$ as a representation of short-term patient conditions.

3) Memory Network: To overcome the shortcoming of RNNs in capturing patient health conditions over the long term, we design a memory network that can 1) preserve fine-grained information of long-term health conditions and 2) provide meaningful fine-grained interpretations. How to refine the memory network to enhance the mutual effects between patient conditions and demographics will be introduced in Section III-C.

Modeling fine-grained patient conditions with key-value memory networks. To model fine-grained patient conditions, we adopt a key-value memory network (KV-MN), which memorizes information by using a large array of external memory slots. The external memories enrich the representation capability compared with hidden vectors of RNNs and enable the KV-MN to capture long-term data characteristics [19]. We aim to fully utilize the representation power and interpretability of KV-MN by carefully designing the memory slots. To achieve this goal, we incorporate the knowledge contained in the disease taxonomy into the memory slots and design each memory slot so that it memorizes patient health conditions on a specific disease category. Compared with RNNs that capture the overall health conditions of a patient, the KV-MN decomposes patient conditions into different

disease categories and thus preserves more fine-grained information. In KV-MNs, a memory slot is represented by a key vector and an associated value vector. Next, we introduce our design of the key vectors, the value vectors, and READ/WRITE operations used to manipulate the memory.

Key vectors. We set the key vectors as the embeddings of the top-level nodes in taxonomy $\mathcal{G}$. This ensures that each memory slot corresponds to a disease category. In particular, the i-th key vector $\mathbf{k}_i \in \mathbb{R}^{d_1}$ is set to the i-th column of the category embedding matrix $\mathbf{K}$. Since $\mathbf{K}$ is computed by using graph-based attention (Section III-B1), it captures the hierarchical information of the taxonomy, e.g., relationships between different diseases. $\mathbf{K}$ is shared by all patients and fixed during the processing of diagnosis sequences.

Value vectors. Let $\mathbf{v}_i$ denote the value vector associated with $\mathbf{k}_i$. Each value vector $\mathbf{v}_i$ memorizes information about patient conditions on one disease category, which helps predict future diagnosis regarding this category. We form a value memory matrix $\mathbf{V} \in \mathbb{R}^{d_v \times L}$ by combining all L value slots. Different from $\mathbf{K}$, $\mathbf{V}$ is patient-specific and is continuously updated according to the input diagnosis sequence. In this way, we capture the dynamic patient conditions on each disease category. Two types of operations, READ and WRITE, are designed to manipulate the value vector.

READ operation. With fine-grained information of historical diagnosis stored in $\{(\mathbf{k}_1, \mathbf{v}_1), \cdots, (\mathbf{k}_L, \mathbf{v}_L)\}$, we obtain long-term patient health conditions from these slots by using the READ operation. Since the patients do not equally suffer from all categories of diseases, we use the short-term representation $\mathbf{h}_t$ as a query to attentively visit the memory network. The attention weight $a_{t,i}$ of $(\mathbf{k}_i, \mathbf{v}_i)$ is calculated according to the correlation between $\mathbf{h}_t$ and $\mathbf{k}_i$:

$$a_{t,i} = \frac{\exp(\mathbf{k}_i^\top \text{MLP}(\mathbf{h}_t))}{\sum_{j=1}^{L} \exp(\mathbf{k}_j^\top \text{MLP}(\mathbf{h}_t))}, \tag{2}$$

where $\text{MLP}(\cdot)$ is a transformation layer composed with a transformation matrix $\mathbf{W}_{tran}$ and a bias vector $\mathbf{b}_{tran}$: $\text{MLP}(\mathbf{h}_t) = \mathbf{W}_{tran}\mathbf{h}_t + \mathbf{b}_{tran}$. Larger $a_{t,i}$ suggests that there is a larger probability that the patient suffers from diseases from the i-th category. The long-term patient health conditions $\mathbf{m}_t$ can thus be represented as an attentive combination of value vectors:

$$\mathbf{m}_t = \sum_{i=1}^{L} a_{t,i}\mathbf{v}_i. \tag{3}$$

WRITE operation. To memorize information of recent diagnosis in the memory network, we update the value matrix $\mathbf{V}$ according to the short-term representation $\mathbf{h}_t$. Inspired by [20], we employ an *erase*-followed-by-*add* update mechanism. This mechanism allows us to erase unnecessary information in the memory and add new information with respect to patient health conditions dynamically.

We first derive an erase vector and an add vector from $\mathbf{h}_t$:

$$\begin{aligned} \mathbf{erase}_t &= \text{sigmoid}(\mathbf{W}_1\mathbf{h}_t + \mathbf{b}_1), \\ \mathbf{add}_t &= \tanh(\mathbf{W}_2\mathbf{h}_t + \mathbf{b}_2), \end{aligned} \tag{4}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_v \times d_2}$, $\mathbf{b}_1 \in \mathbb{R}^{d_v}$, $\mathbf{W}_2 \in \mathbb{R}^{d_v \times d_2}$, and $\mathbf{b}_2 \in \mathbb{R}^{d_v}$ are parameters of the erase layer and the add layer. Here sigmoid$(\cdot)$ and tanh$(\cdot)$ are chosen as the activation functions of the erase layer and the add layer following [20]. Since memory slots that are associated with patient health conditions should be emphasized during the update, the WRITE operation is performed attentively by considering the attention weight $a_{t,i}$:

$$\mathbf{v}_i \leftarrow \mathbf{v}_i \odot (\mathbf{1} - a_{t,i}\mathbf{erase}_t) + a_{t,i}\mathbf{add}_t, \tag{5}$$

where $\odot$ is the Hadamard product and $\mathbf{1}$ is a d_v-dimensional column vector of all 1's. By learning the parameters of the erase and add layers, our model can automatically determine which signals to weaken or strengthen based on recent diagnosis.

C. Co-Attention-Based Mutual Enhancement

To effectively model the mutual effect between patient demographics and the memory network, we design a co-attention mechanism that consists of **attention for memory slots** and **attention for patient demographics**. In this way, CAMP can accurately predict future diagnosis with mutually enhanced representations of long-term memory and patient demographics.

1) Attention for Memory Slots: It often happens that patients with certain demographics are vulnerable to some diseases while others are not. For instance, HFMD (hand, foot, and mouth disease) typically occurs in children instead of adults [21]. It inspires us to consider patient demographics $\mathbf{p}$ when computing the attention weight $a_{t,i}$ of the i-th disease category. Specifically, we first obtain the demographics embedding $\mathbf{q} \in \mathbb{R}^{d_3}$ from the raw multi-hot vector $\mathbf{p} \in \{0, 1\}^r$ via an embedding layer:

$$\mathbf{q} = \mathbf{W}_p\mathbf{p} + \mathbf{b}_p, \tag{6}$$

where $\mathbf{W}_p \in \mathbb{R}^{d_3 \times r}$ and $\mathbf{b}_p \in \mathbb{R}^{d_3}$ are parameters of the embedding layer and d_3 is the embedding size. Then, we use the concatenation of $\mathbf{h}_t$ and $\mathbf{q}$ as the query to visit the memory network. Thus, the calculation of $a_{t,i}$ in Equation (2) is replaced with:

$$a_{t,i} = \frac{\exp(\mathbf{k}_i^\top \text{MLP}(\mathbf{h}_t \oplus \mathbf{q}))}{\sum_{j=1}^{L} \exp(\mathbf{k}_j^\top \text{MLP}(\mathbf{h}_t \oplus \mathbf{q}))}, \tag{7}$$

where $\oplus$ is the concatenation operator. We can manipulate the memory network better with the enhanced memory attention mechanism that takes patient demographics into consideration.

2) Attention for Patient Demographics: Given a patient with the demographics embedding $\mathbf{q}$, the long-term memory representation $\mathbf{m}_t$ can serve as an important context about historical diagnosis and help decide which latent features in $\mathbf{q}$ are more important for prediction. Thus, we leverage $\mathbf{m}_t$ to enhance the original representation $\mathbf{q}$ and derive an attention vector $\boldsymbol{\beta} \in \mathbb{R}^{d_3}$ for $\mathbf{q}$ as

$$\boldsymbol{\beta} = \text{ReLU}(\mathbf{W}_3\mathbf{m}_t + \mathbf{W}_4\mathbf{q} + \mathbf{b}_3), \tag{8}$$

where $\mathbf{W}_3 \in \mathbb{R}^{d_3 \times d_v}$, $\mathbf{W}_4 \in \mathbf{R}^{d_3 \times d_3}$, and $\mathbf{b}_3 \in \mathbb{R}^{d_3}$ are parameters of the attention layer. Conditioned on the historical

diagnosis, $\boldsymbol{\beta}$ is used to enhance the original representation of demographics:

$$\mathbf{q}^{new} = \boldsymbol{\beta} \odot \mathbf{q}. \tag{9}$$

D. Joint Learning

Given the short-term representation $\mathbf{h}_t$, the long-term representation $\mathbf{m}_t$ of patient health conditions and the enhanced representation of patient demographics $\mathbf{q}^{new}$, we generate a joint representation of patient by concatenating the three representations. The concatenated vector is fed through a softmax layer to predict the diagnosis of next visit $\tilde{\mathbf{x}}_{t+1}$:

$$\tilde{\mathbf{x}}_{t+1} = \text{softmax}(\mathbf{W}_x(\mathbf{h}_t \oplus \mathbf{m}_t \oplus \mathbf{q}^{new}) + \mathbf{b}_x). \tag{10}$$

Here $\mathbf{W}_x \in \mathbb{R}^{|\mathcal{C}|\times(d_2+d_v+d_3)}$ and $\mathbf{b}_x \in \mathbb{R}^{|\mathcal{C}|}$ are parameters to be learned. Following [8], [7], we use the cross-entropy between the ground truth $\mathbf{x}_{t+1}$ and the predicted $\tilde{\mathbf{x}}_{t+1}$ to calculate the loss for each patient:

$$\mathcal{L} = -\frac{1}{T-1}\sum_{t=1}^{T-1}(\mathbf{x}_{t+1}^{\top}\log(\tilde{\mathbf{x}}_{t+1}) + (\mathbf{1}-\mathbf{x}_{t+1})^{\top}\log(\mathbf{1}-\tilde{\mathbf{x}}_{t+1})). \tag{11}$$

The loss of all patients can be calculated by averaging $\mathcal{L}$. Note that all parameters in the neural architecture can be jointly optimized in an end-to-end way. We use the Adam optimizer [22] because it can automatically adjust the learning rate during the training phase.

IV. Experiments

In this section, we conduct experiments to evaluate our approach. We aim to answer the following research questions:

- **RQ1:** How does CAMP perform compared with state-of-the-art diagnosis prediction models?
- **RQ2:** How do different components (i.e., the external memories, the GRU, the demographics embeddings, and co-attention-based mutual enhancement) affect CAMP?
- **RQ3:** Can CAMP provide meaningful interpretations on progression patterns of patient conditions and properly relate them with patient demographics?

A. Experimental Settings

Table I: Statistics of two datasets.

Dataset	DPH	MIMIC-III
# of patients	46,074	7,499
# of visits	447,505	19,911
Avg. # of visits per patient	9.71	2.66
# of unique ICD codes	6,059	4,880
Avg. # of ICD codes per visit	2.42	13.06
Max. # of ICD codes per visit	27	39
# of unique CCS group codes	238	272
Avg. # of CCS group codes per visit	2.32	11.23
Max. # of CCS group codes per visit	24	34
# of top-level codes	17	17

1) Datasets: We conduct experiments on two real-world EHR datasets: the DPH dataset and the MIMIC-III dataset.

- **DPH Dataset** consists of medical records of 46,074 patients collected by Peking University Peoples Hospital from 2009 to 2014. Following [4], we filter out sequences that are too short in length. Only patients with at least 5 visits are preserved in the dataset. This dataset helps evaluate how diagnosis prediction methods perform on long diagnosis sequences.
- **MIMIC-III Dataset**[4] is a publicly available EHR dataset containing medical records of 7,499 intensive care unit (ICU) patients over 11 years. Following [8], we only choose patients with at least two visits. Since MIMIC-III consists of very short visits and the number of patients is small, it helps evaluate the performance of prediction approaches on high-risk patients with insufficient training data.

The diagnosed diseases are represented with ICD-10 codes in the DPH dataset and with ICD-9 codes in the MIMIC-III dataset. To improve the training speed and preserve sufficient granularity of each diagnosis, we group the ICD codes by using CCS single-level diagnosis grouper[5] and replace the original ICD codes with their group codes following [7]. We use CCS-multi-level diagnosis hierarchy[6] as the taxonomy of diseases. Since there are 18 top-level codes in CCS multi-level hierarchy and the last one represents residual and unclassified disease codes, we only use the first 17 ones on both datasets. Detailed statistics of two datasets are shown in Table I.

2) Models for comparison: We select six competitive models for comparison, which can be divided into three groups.

G1: Models that utilize only historical records. Models in G1 handle diagnosis sequences without incorporating auxiliary information such as taxonomies of diseases and demographics.

- **RNN**. The diagnosis $\mathbf{x}_t$ of t-th visit is embedded into $\mathbf{e}_t$ and fed into the GRU module. It directly predicts future diagnosis based on the hidden state vector $\mathbf{h}_t$.
- **RNN+**. It combines hidden state vectors of previous visits by adding location-based attention mechanism [6] into RNNs.
- **Dipole** [6]. It replaces the vanilla RNN with a bidirectional one to utilize available information in the past and future.

Timeline [23] is not a baseline as it requires additional temporal information (e.g., time of next visit) to predict future clinical events.

G2: Demographics-aware model. The model in G2 utilizes patient demographics in diagnosis prediction.

- **MCA-RNN** [4]. This is a hybrid model that utilizes an attention-based RNN and a conditional variational autoencoder to capture information in patient demographics.

G3: Taxonomy-aware models. Models in G3 incorporate taxonomies of diseases for diagnosis prediction.

- **GRAM** [7]. With the taxonomy of diseases, GRAM learns robust and reasonable embeddings of diseases

[4]https://mimic.physionet.org/

[5]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixASingleDX.txt

[6]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixCMultiDX.txt

Table II: Performance of models on two datasets. Best results are highlighted in bold. Models in G1 consider only historical records, G2 represents demographics-aware models, G3 denotes taxonomy-aware models. The symbol * means that the improvement is significant with p-value < 0.001 by t-test.

Dataset	Group	Method	Recall@K			MAP@K		
			K=5	K=10	K=15	K=5	K=10	K=15
DPH	G1	RNN	0.672±0.002	0.769±0.002	0.814±0.001	0.547±0.001	0.577±0.001	0.585±0.001
		RNN+	0.667±0.001	0.765±0.002	0.811±0.001	0.542±0.001	0.572±0.001	0.581±0.001
		Dipole	0.665±0.002	0.756±0.002	0.799±0.002	0.542±0.003	0.570±0.002	0.578±0.002
	G2	MCA-RNN	0.677±0.002	0.783±0.001	0.824±0.002	0.549±0.002	0.581±0.001	0.588±0.002
	G3	GRAM	0.679±0.002	0.778±0.001	0.822±0.001	0.554±0.001	0.583±0.001	0.591±0.001
		KAME	0.680±0.002	0.777±0.001	0.821±0.001	0.556±0.002	0.585±0.002	0.593±0.002
	Ours	CAMP	**0.694±0.001***	**0.791±0.001***	**0.833±0.001***	**0.567±0.002***	**0.596±0.001***	**0.604±0.001***
MIMIC-III	G1	RNN	0.288±0.002	0.432±0.002	0.528±0.002	0.245±0.002	0.333±0.001	0.380±0.002
		RNN+	0.289±0.002	0.431±0.003	0.527±0.003	0.247±0.002	0.334±0.002	0.381±0.003
		Dipole	0.282±0.002	0.423±0.002	0.520±0.002	0.239±0.002	0.323±0.002	0.370±0.002
	G2	MCA-RNN	0.291±0.001	0.438±0.001	**0.539±0.001**	0.248±0.002	0.340±0.002	0.391±0.001
	G3	GRAM	0.293±0.002	0.437±0.002	0.535±0.003	0.252±0.002	0.341±0.002	0.389±0.002
		KAME	0.292±0.002	0.438±0.002	0.535±0.002	0.249±0.002	0.339±0.002	0.387±0.003
	Ours	CAMP	**0.297±0.001***	**0.443±0.001***	**0.539±0.002**	**0.256±0.001***	**0.347±0.001***	**0.396±0.001***

with a graph-based attention mechanism for performance improvement.

- **KAME** [8]. This model explicitly makes use of medical knowledge in the whole prediction process.

3) Evaluation Criteria: To comprehensively evaluate the performance of models, we use the following two criteria:

- **Recall@**K is defined as the number of correct codes in top K of $\tilde{\mathbf{x}}_{t+1}$ divided by the number of all correct codes, which is widely used by other studies on diagnosis prediction [4], [5].
- **MAP@**K (mean average precision) is a widely used metric in information retrieval [24], [25]. We use this metric to consider the orders of correctly predicted codes.

We vary K in $\{5, 10, 15\}$ for a more thorough evaluation.

4) Implementation Details: We treat visits of each patient as a sample and randomly split the dataset into training (75%), validation (10%) and testing (15%) sets as [8]. We report performance according to predictions for the last visit of patients in the testing set. For fair consideration, all models are optimized using Adam [22] with an initial learning rate of 0.001 and the batch size is fixed to 100. The coefficient of L_2 norm regularization is fixed to 0.001. The size r of patient demographics vector is 7 (2 genders + 5 age groups) in the DPH dataset and 11 (2 genders + 5 age groups + 4 admission types) in the MIMIC-III dataset. We tune hyper-parameters of models on the validation set. In CAMP, we set d_1, d_2, d_v to 128, 256, and 48 for the DPH dataset. We set d_1, d_2, d_v to 128, 384, and 16 for the MIMIC-III dataset. Following [20], we learn the initial value of $\mathbf{V}$ in the training process, which represents the initial health condition with respect to each disease category. Each experiment is repeated ten times and we report the average and standard deviation as the result. To ensure reproducibility, detailed instructions on running our model has been provided along with the source code.

B. Performance Comparison (RQ1)

The diagnosis prediction results of CAMP and all six baselines on two datasets are given in Table II. We further conduct paired t-tests showing whether the improvements of CAMP are statistically significant (e.g., p-value < 0.001). Four observations are made from Table II.

First, our model CAMP outperforms all state-of-the-art models. On the DPH dataset, CAMP achieves 2.1% higher Recall@5 and 2.0% higher MAP@5 over all baselines. Moreover, the improvements in terms of all criteria except Recall@15 on the MIMIC-III dataset are statistically significant. This demonstrates the effectiveness of our proposed framework of co-attention memory networks, which allows CAMP to capture complicated patient health conditions from diagnosis sequences. The superiority of CAMP also stems from its design that jointly models the mutual effect between important context (i.e., medical knowledge and patient demographics) and historical records while baselines fail to do so. We also observe that the overall improvement of CAMP on the DPH dataset is more significant than that on the MIMIC-III dataset. This is ascribed to the fact that the average length of diagnosis sequences on the former is much longer than the latter, which makes it easier to highlight the strength of memory networks in handling long sequences. Even for the MIMIC-III dataset, which consists of short sequences (on average 2.66 visits for one patient), our method is able to achieve stable accuracy gain compared with the baselines.

Second, demographics-aware model (MCA-RNN) performs better than the models that consider only historical records (RNN, RNN+ and Dipole), achieving 1.4% higher Recall@K

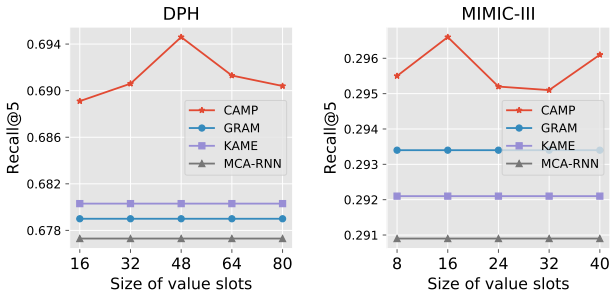

Figure 2: Recall@5 of CAMP and baselines on two datasets with different sizes of memory value slots (d_v).

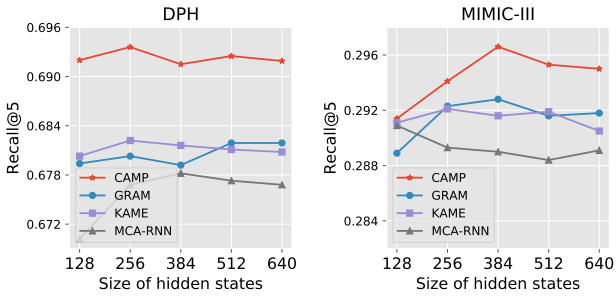

Figure 3: Recall@5 of CAMP and baselines on two datasets with different sizes of GRU hidden states (d_2).

and 1.6% higher Map@K on the MIMIC-III dataset. It demonstrates that patient demographics are important contextual information for modeling the diagnosis sequences and help improve the prediction performance.

Third, taxonomy-aware models (GRAM and KAME) generally achieve better performance than the models that consider only historical records. The mean improvements of Recall@K on two datasets are 1.1% and 1.4%. The mean improvements of MAP@K on two datasets are 1.5% and 2.1% respectively. We ascribe these improvements to the fact that they learn better disease embeddings that capture intrinsic characteristics with medical knowledge and thus predict future diagnosis more accurately.

Fourth, the performance of Dipole is worse than that of RNN on the two datasets, which is consistent with results reported in [8]. This indicates that the bi-directional mechanism in Dipole sometimes introduces noises to the representation of the last visit, which is important in prediction. This may be a reason why state-of-the-art models such as MCA-RNN, KAME and GRAM are based on the vanilla RNN rather than a bi-directional one.

C. Detailed Analysis of CAMP (RQ2)

As shown above, our model CAMP has achieved a significant improvement over all the baselines. In this section, we conduct detailed analysis of CAMP to better understand the influence of its different components.

Specifically, we study the effect of the external memories, the GRU, and the demographics embeddings by varying three

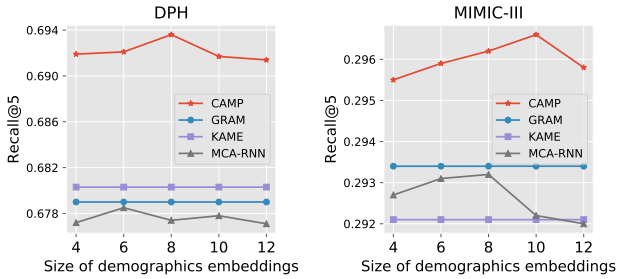

Figure 4: Recall@5 of CAMP and baselines on two datasets with different sizes of demographics embeddings (d_3).

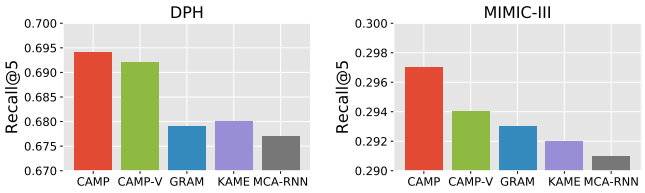

Figure 5: Recall@5 of CAMP, CAMP-V (the variant without co-attention-based mutual enhancement) and baselines on two datasets.

corresponding hyper-parameters of our model, including (1) the size of memory value slots d_v; (2) the size of GRU hidden states d_2; and (3) the size of demographics embeddings d_3. When conducting analysis for one hyper-parameter, we set other hyper-parameters to values described in Section IV-A4. We further study the effect of co-attention-based mutual enhancement by comparing CAMP with one variant.

Figures 2-4 show how prediction performance changes with different hyper-parameters. Figure 5 shows the prediction performance with and without the mechanism of co-attention-based mutual enhancement. Three most competitive baselines (GRAM, KAME and MCA-RNN) are selected for comparison. Due to space limitations, we only show the results of Recall@5 on two datasets. The results regarding other evaluation criteria are similar. Based on the observation, we draw four conclusions.

Robustness of our method. CAMP consistently outperforms three most competitive baselines with varying sizes of memory value slots, hidden states, and demographics embeddings. This demonstrates the robustness of our method. Note that d_v is not a hyper-parameter for the three baselines, and d_3 is not a hyper-parameter for GRAM and KAME. Thus, their performance remains the same when we change the corresponding hyper-parameters.

Effectiveness of the external memories. Figure 2 shows that our method achieves the best performance on DPH (or MIMIC-III) when d_v is equal to 48 (or 16). Smaller d_v leads to insufficient representation of fine-grained patient conditions and larger d_v may result in over fitting. This demonstrates the importance of using external memories, which contain fine-grained information about long-term patient health conditions regarding each category of diseases.

Effectiveness of the GRU and the demographics embeddings. Figures 3 and 4 show that too small or too large values of d_2 and d_3 will hurt the prediction performance of CAMP. This illustrates the importance of properly incorporating short-term information learned by the GRU and encoding patient demographics.

Effectiveness of co-attention-based mutual enhancement To validate the effectiveness of the co-attention mechanism in our model, we compare CAMP with one variant: CAMP-V. CAMP-V disables the interaction between patient demographics and the memory network. Specifically, in CAMP-V, $\mathbf{q}$ (instead of $\mathbf{q}^{\mathbf{new}}$) is utilized in the final prediction layer. The memory attention is calculated by using Equation (2) instead of Equation (7). As shown in Figure 5, CAMP performs better than CAMP-V on two datasets, which confirms our assumption that modeling the mutual effects between long-term patient conditions and demographics results in improved prediction performance.

D. Case Study on Model Interpretability (RQ3)

In the previous experiments, we have demonstrated that CAMP enhances the accuracy of diagnosis prediction. Another major benefit is that the model is highly interpretable due to the co-attention mechanism and the incorporation of the disease taxonomy in the KV-MN. To illustrate the interpretability of CAMP, we conduct a case study with an experienced clinical expert. By investigating the interpretations provided by CAMP, the expert is able to understand 1) how the model captures fine-grained progression patterns of a patient and 2) how it connects the health conditions of a patient with his/her demographics.

Understanding fine-grained progression patterns of a patient. The memory attentions at different visits enable us to understand fine-grained progression patterns of a patient. Case I in Table III shows four visits of a 57-year-old female patient (from DPH) and the related interpretations. This patient is chosen because her visit sequence is the longest (20 visits). Names of diseases are presented along with the corresponding CCS codes that indicate their positions in the taxonomy. Only the disease categories (memory slots) with large attention values are displayed.

First, we can observe that CAMP correctly captures **long-term fine-grained** (category-level) patient conditions. In most of her visits, the patient is diagnosed with diseases (3.6) and (7.2), which are *disorders of lipid metabolism* and *essential hypertension*. Our model captures this pattern and learns to assign largest attention weights on memory slots that correspond to disease categories (3) and (7). Moreover, the attention weight of category (9), *diseases of digestive systems*, is relatively large, although the patient has never been diagnosed with diseases belonging to this category. This is confirmed positively by the clinical expert and related literature since some *disorders of lipid metabolism* (e.g., *hyperlipidemia*) are a known risk of *fatty infiltration of the liver*, which is one of *diseases of digestive systems* [26]. Thus, it is reasonable for CAMP to consider that the patient may suffer from disease in category (9) (*diseases of digestive systems*) even without such historical diagnosis. Because CAMP correctly captures fine-grained long-term conditions of the patient, it is able to correctly predict diagnosed diseases of the $(t+1)$-th visit even if some diseases have never been diagnosed. For example, CAMP correctly predicts disease (7.2.2), which has never been diagnosed and is a complication of *essential hypertension* [27].

Second, CAMP is able to capture the **dynamic changes** in patient health conditions. In visit t-3, the patient was diagnosed with disease (6.7.6). Our model captures this change and increases the attention weight of category (6). Since the disease has never been diagnosed after visit t-3, the attention weight of category (6) becomes smaller after the visit.

Connecting patient conditions with demographics. In CAMP, different features in $\mathbf{q}$ are strengthened or weakened via the attention vector $\boldsymbol{\beta}$ to derive an enhanced demographics representation $\mathbf{q}^{new}$. This allows us to interpret the importance of raw demographics $\mathbf{p}$. According to Choi et al. [3], we can calculate the contribution-coefficient of each raw feature $\mathbf{p}_i$ in $\mathbf{q}^{new}$ as $\boldsymbol{\beta}^\top \mathbf{W}_p[:, i]$, where $\mathbf{W}_p[:, i]$ is the i-th column of $\mathbf{W}_p$. In this case study, we adopt a 7-dimensional vector to represent the patient. The first two dimensions correspond to two genders and the other five dimensions stand for five age groups ($\leq$18, 18-40, 40-50, 50-80, $>$80).

We find that our model is capable of enhancing the most important features based on patient conditions. For the patient in Case I, the contribution-coefficients of feature *Age>80* and feature *50<Age≤80* are significantly larger than others. This indicates that CAMP mainly cares whether the patient belongs to these age groups when predicting future diagnosis related to *hypertension*. The expert says that this interpretation is supported by the medical literature [28], which claims that increasing age is an independent risk factor for the development of diseases related to *hypertension* (e.g., *atherosclerosis*). In comparison, for the 38-year-old female patient in DPH (Case-II, Table III), the gender of female contributes most to her demographics representation. This is reasonable since the patient has been diagnosed with diseases (10.3.6) and (10.3.9), which are diseases corresponding to the female genital system.

V. Related Work

In this section, we review the studies related to our work.

A. Diagnosis Prediction

The broad adoption of EHR systems has opened the possibility of gaining knowledge by mining massive EHR data [29], [30]. Typical tasks of EHR data mining include adverse drug reaction detection [31], [32], phenotyping [10], patient subtyping [33], [34], disease progression [35], [36], and diagnosis prediction [3], [4], [5], [6], [7], [8], [18].

Diagnosis prediction, which aims to predict future diagnosis according to historical visit information of patients, is one important task in EHR data mining. DoctorAI [5] uses a two-layer RNN model to predict future diagnosis and visit time. RETAIN [3] is an interpretable predictive model using a reverse time attention mechanism in RNNs. Dipole [6] replaces

Table III: Case study on interpretability of CAMP.

Case I		
Visit t-4	Diagnosed Diseases	Disorders of lipid metabolism (3.6); Essential hypertension (7.1.1)
	Demographics coefficients	**Age>80**: 1.312; **50<Age≤80**: 0.275; Female: 0.090
	Memory attentions	Endocrine; nutritional; and metabolic diseases and immunity disorders (3): 0.489; Diseases of the circulatory system (7): 0.043; **Diseases of the digestive system (9)**: 0.041
Visit t-3	Diagnosed Diseases	Other eye disorders (6.7.6)
	Demographics coefficients	**Age>80**: 1.448; **50<Age≤80**: 0.305; Female: 0.122
	Memory attentions	Endocrine; nutritional; and metabolic diseases and immunity disorders (3): 0.441; Diseases of the circulatory system (7): 0.152; **Diseases of the nervous system and sense organs (6)**: 0.025
Visit t	Diagnosed Diseases	Disorders of lipid metabolism (3.6); Essential hypertension (7.1.1)
	Demographics coefficients	**Age>80**: 1.439; **50<Age≤80**: 0.326; Female: 0.126
	Memory attentions	Endocrine; nutritional; and metabolic diseases and immunity disorders (3): 0.515; Diseases of the circulatory system (7): 0.057; **Diseases of the digestive system (9)**: 0.035
Visit t+1	Diagnosed Diseases	Disorders of lipid metabolism (3.6); Essential hypertension (7.1.1); **Coronary atherosclerosis and other heart disease (7.2.2)**
Case II		
Visit t	Diagnosed Diseases	Ovarian cyst (10.3.6); Other female genital disorders (10.3.9); Deficiency and other anemia (4.1.3)
	Demographics coefficients	**Female**: 0.420; 18<Age≤40: 0.205; 50<Age≤80: 0.031
	Memory attentions	Diseases of the blood and blood-forming organs (4): 0.594; Diseases of the respiratory system (8): 0.070; Diseases of the genitourinary system (10): 0.027

the vanilla RNN with a bi-directional one for further improvement. Baytas et al. [34] and Bai et al. [23] propose to model the elapsed time between consecutive visits in RNNs. Xiao et al. adopt a hybrid model TopicRNN that combines topic models with RNNs [18]. Recently, some pioneer studies try to incorporate auxiliary information to further improve prediction accuracy. MCA-RNN [4] leverages a conditional variational auto-encoder (CVAE) [37] that takes patient demographics as contextual information. Choi et al. [7] incorporate taxonomies of diseases to train better embeddings for diagnosis codes and Ma et al. [8] make explicit use of such medical knowledge in KAME for performance improvement. However, due to the limited representation power of hidden vectors in RNNs, these methods cannot model fine-grained information of diagnosed diseases and long sequences effectively.

Compared with existing methods, our proposed CAMP is a novel memory-augmented model which can better capture patient health conditions during diagnosis sequences. The design of KV-MN and co-attention mechanism also allows it to take full advantage of patient demographics and medical knowledge in prediction. Moreover, our method is capable of providing meaningful fine-grained interpretations about dynamic patient conditions and connecting them with important context (e.g., age and gender of the patients).

B. Memory Networks

Memory networks emerge recently as a powerful framework to process sequential data. The memory component increases model capacity and enables the neural network to track long-term dependencies. The initial framework of memory networks is proposed by Weston et al. [9]. Following this idea, Sukhbaatar et al. propose an end-to-end memory-augmented model that significantly reduces the requirement of supervision during training [38]. Miller et al. propose KV-MNs that decompose the memory component into the key part and the value part [13]. Due to its superiority in sequence modeling, researchers have applied memory networks in sequential prediction tasks such as question answering [14], natural language transduction [39], knowledge tracing [20], asynchronous multi-view learning [40] and sequential recommendation [15], [41]. In this paper, we demonstrate how memory networks can be used in diagnosis prediction and cooperate with patient demographics in a mutual enhancement manner, which has not been explored in the research community. Our design of the memory networks enables accuracy improvement over existing methods in diagnosis prediction and ensures that the model is highly interpretable.

VI. Conclusions

In this paper, we propose a model named co-attention memory networks (CAMP) for diagnosis prediction. The model adopts a three-way interaction architecture to tightly integrate historical records, fine-grained patient conditions, and demographics. The analysis of fine-grained patient conditions is enabled by explicitly incorporating taxonomies of diseases into a memory network. We elaborately design the memory network to ensure that it provides meaningful interpretations and cooperates with patient demographics in a mutual enhancement manner. Experiments and a case study on real-world datasets demonstrate that CAMP consistently performs better than state-of-the-art methods in terms of prediction accuracy and is highly interpretable.

VII. ACKNOWLEDGMENT

This work is supported by the National Science and Technology Major Project (No. 2018ZX10201002).

REFERENCES

[1] J. Zhou, J. Sun, Y. Liu, J. Hu, and J. Ye, "Patient risk prediction model via top-k stability selection," in *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 2013, pp. 55–63.

[2] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, "Risk prediction on electronic health records with prior medical knowledge," in *KDD*. ACM, 2018, pp. 1910–1919.

[3] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, 2016, pp. 3504–3512.

[4] W. Lee, S. Park, W. Joo, and I.-C. Moon, "Diagnosis prediction via medical context attention networks using deep generative modeling," in *ICDM*. IEEE, 2018, pp. 1104–1109.

[5] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference*, 2016, pp. 301–318.

[6] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *KDD*. ACM, 2017, pp. 1903–1911.

[7] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "Gram: graph-based attention model for healthcare representation learning," in *KDD*. ACM, 2017, pp. 787–795.

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

[9] J. Weston, S. Chopra, and A. Bordes, "Memory networks," in *ICLR*, 2015.

[10] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *AAAI*, 2018, pp. 4091–4098.

[11] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *KDD*. ACM, 2016, pp. 1495–1504.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *KDD*. ACM, 2016, pp. 1135–1144.

[13] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," *arXiv preprint arXiv:1606.03126*, 2016.

[14] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *ICML*, 2016, pp. 1378–1387.

[15] X. Chen, H. Xu, Y. Zhang, J. Tang, Y. Cao, Z. Qin, and H. Zha, "Sequential recommendation with user memory networks," in *WSDM*. ACM, 2018, pp. 108–116.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] K. Cho, B. Van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *Computer Science*, 2014.

[18] C. Xiao, T. Ma, A. B. Dieng, D. M. Blei, and F. Wang, "Readmission prediction via deep contextual embedding of clinical concepts," *PloS one*, vol. 13, no. 4, p. e0195024, 2018.

[19] F. Liu and J. Perez, "Gated end-to-end memory networks," in *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1, 2017, pp. 1–10.

[20] J. Zhang, X. Shi, I. King, and D.-Y. Yeung, "Dynamic key-value memory networks for knowledge tracing," in *WWW*. International World Wide Web Conferences Steering Committee, 2017, pp. 765–774.

[21] L.-Y. Chang, T.-Y. Lin *et al.*, "Clinical features and risk factors of pulmonary oedema after enterovirus-71-related hand, foot, and mouth disease," *The Lancet*, vol. 354, no. 9191, pp. 1682–1686, 1999.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] T. Bai, S. Zhang, B. L. Egleston, and S. Vucetic, "Interpretable representation learning for healthcare via capturing disease progression through time," in *KDD*. ACM, 2018, pp. 43–51.

[24] F. Raiber and O. Kurland, "Ranking document clusters using markov random fields," in *SIGIR*. ACM, 2013, pp. 333–342.

[25] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W.-Y. Ma, "Collaborative knowledge base embedding for recommender systems," in *KDD*. ACM, 2016, pp. 353–362.

[26] N. Assy, K. Kaita, D. Mymin, C. Levy, B. Rosser, and G. Minuk, "Fatty infiltration of liver in hyperlipidemic patients," *Digestive diseases and sciences*, vol. 45, no. 10, pp. 1929–1934, 2000.

[27] V. J. Dzau, "Atherosclerosis and hypertension: mechanisms and interrelationships." *Journal of cardiovascular pharmacology*, vol. 15, pp. S59–64, 1990.

[28] J. C. Wang and M. Bennett, "Aging and atherosclerosis: mechanisms, functional consequences, and potential therapeutics for cellular senescence," *Circulation research*, vol. 111, no. 2, pp. 245–259, 2012.

[29] A. D. Black, J. Car, C. Pagliari, C. Anandan, K. Cresswell, T. Bokun, B. McKinstry, R. Procter, A. Majeed, and A. Sheikh, "The impact of ehealth on the quality and safety of health care: a systematic overview," *PLoS medicine*, vol. 8, no. 1, p. e1000387, 2011.

[30] A. K. Jha, C. M. DesRoches, E. G. Campbell, K. Donelan, S. R. Rao, T. G. Ferris, A. Shields, S. Rosenbaum, and D. Blumenthal, "Use of electronic health records in us hospitals," *New England Journal of Medicine*, vol. 360, no. 16, pp. 1628–1638, 2009.

[31] C. Xiao, P. Zhang, W. A. Chaowalitwongse, J. Hu, and F. Wang, "Adverse drug reaction prediction with symbolic latent dirichlet allocation," in *AAAI*, 2017.

[32] B. Jin, H. Yang, C. Xiao, P. Zhang, X. Wei, and F. Wang, "Multitask dyadic prediction and its application in prediction of adverse drug-drug interaction," in *AAAI*, 2017.

[33] Z. Zhu, C. Yin, B. Qian, Y. Cheng, J. Wei, and F. Wang, "Measuring patient similarities via a deep architecture with medical concept embedding," in *ICDM*. IEEE, 2016, pp. 749–758.

[34] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware lstm networks," in *KDD*. ACM, 2017, pp. 65–74.

[35] H. Xiao, J. Gao, L. Vu, and D. S. Turaga, "Learning temporal state of diabetes patients via combining behavioral and demographic data," in *KDD*. ACM, 2017, pp. 2081–2089.

[36] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *KDD*. ACM, 2014, pp. 85–94.

[37] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, 2015, pp. 3483–3491.

[38] S. Sukhbaatar, J. Weston, R. Fergus *et al.*, "End-to-end memory networks," in *Advances in neural information processing systems*, 2015, pp. 2440–2448.

[39] E. Grefenstette, K. M. Hermann, M. Suleyman, and P. Blunsom, "Learning to transduce with unbounded memory," *Neural information processing systems*, pp. 1828–1836, 2015.

[40] H. Le, T. Tran, and S. Venkatesh, "Dual memory neural computer for asynchronous two-view sequential learning," in *KDD*. ACM, 2018, pp. 1637–1645.

[41] J. Huang, W. X. Zhao, H. Dou, J.-R. Wen, and E. Y. Chang, "Improving sequential recommendation with knowledge-enhanced memory networks," in *SIGIR*. ACM, 2018, pp. 505–514.