Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks

Fenglong Ma*
SUNY Buffalo
Buffalo, NY, USA
fenglong@buffalo.edu

Radha Chitta
Conduent Labs US
Rochester, NY, USA
radha.chitta@conduent.com

Jing Zhou
Conduent Labs US
Rochester, NY, USA
jing.zhou@conduent.com

Quanzeng You
University of Rochester
Rochester, NY, USA
quanzeng.you@rochester.edu

Tong Sun†
United Technologies Research Center
East Hartford, CT, USA
sunt@utrc.utc.com

Jing Gao
SUNY Buffalo
Buffalo, NY, USA
jing@buffalo.edu

ABSTRACT

Predicting the future health information of patients from the historical Electronic Health Records (EHR) is a core research task in the development of personalized healthcare. Patient EHR data consist of sequences of visits over time, where each visit contains multiple medical codes, including diagnosis, medication, and procedure codes. The most important challenges for this task are to model the temporality and high dimensionality of sequential EHR data and to interpret the prediction results. Existing work solves this problem by employing recurrent neural networks (RNNs) to model EHR data and utilizing simple attention mechanism to interpret the results. However, RNN-based approaches suffer from the problem that the performance of RNNs drops when the length of sequences is large, and the relationships between subsequent visits are ignored by current RNN-based approaches. To address these issues, we propose Dipole, an end-to-end, simple and robust model for predicting patients' future health information. Dipole employs bidirectional recurrent neural networks to remember all the information of both the past visits and the future visits, and it introduces three attention mechanisms to measure the relationships of different visits for the prediction. With the attention mechanisms, Dipole can interpret the prediction results effectively. Dipole also allows us to interpret the learned medical code representations which are confirmed positively by medical experts. Experimental results on two real world EHR datasets show that the proposed Dipole can significantly improve the prediction accuracy compared with the state-of-the-art diagnosis prediction approaches and provide clinically meaningful interpretation.

CCS CONCEPTS

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

KEYWORDS

Healthcare informatics, bidirectional recurrent neural networks, attention mechanism

ACM Reference format:
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 KDD'17, August 13–17, 2017, Halifax, NS, Canada.*, 9 pages.
DOI: http://dx.doi.org/10.1145/3097983.3098088

1 INTRODUCTION

Electronic Health Records (EHR), consisting of longitudinal patient health data, including demographics, diagnoses, procedures, and medications, have been utilized successfully in several predictive modeling tasks in healthcare [9–11, 31]. EHR data are temporally sequenced by patient medical visits that are represented by a set of high dimensional clinical variables (i.e., medical codes). One critical task is to predict the future diagnoses based on patient's historical EHR data, i.e., *diagnosis prediction*. When predicting diagnoses, each patient's visit and the medical codes in each visit may have varying importance. Thus, the most important and challenging issues in diagnosis prediction are:

- How to correctly model such *temporal* and *high dimensional* EHR data to significantly improve the performance of prediction;
- How to reasonably *interpret* the importance of visits and medical codes in the prediction results.

In order to model sequential EHR data, recurrent neural networks (RNNs) have been employed in the literature for deriving accurate and robust representations of patient visits in diagnosis prediction task [10, 11]. RETAIN [11] and GRAM [10] are two state-of-the-art models utilizing RNNs for predicting the future diagnoses. RETAIN applies an RNN with reverse time ordered EHR sequences, while GRAM uses an RNN when modeling time ordered patient visits. Both models achieve good prediction accuracy. However, they are constrained by the forgetfulness associated with models using RNNs, i.e. RNNs cannot handle long sequences effectively.

*This work was mostly done when the first author was an intern in Xerox.
†This work was done when the fifth author was part of Xerox.

KDD'17, August 13–17, 2017, Halifax, NS, Canada.

DOI: http://dx.doi.org/10.1145/3097983.3098088

The predictive power of these models drops significantly when the length of the patient visit sequences is large. Bidirectional recurrent neural networks (BRNNs) [24], which can be trained using all available input information in the past and future, have been used to alleviate the effect of the long sequence problem, and improve the predictive performance.

However, it is infeasible to interpret the outputs of models incorporating either RNNs or BRNNs. Interpretability is crucial in the healthcare domain, as it can lead to the identification of potential risk factors and the design of suitable intervention mechanisms. Non-temporal models such as Med2Vec [9] generate easily interpretable low-dimensional representations of the medical codes, but do not account for the temporal nature of the EHR data.

To model the temporal EHR data and interpret the prediction results simultaneously, attention-based neural networks can be applied, which aim to learn the relevance of the data samples to the task. For example, RETAIN [11] employs location-based attention to predict the future diagnosis. It calculates the attention weights for a visit at time t, using the medical information in the current visit and the hidden state of the recurrent neural network at time t, to predict the visit at time $t+1$. It ignores the relationships between all the visits from time 1 to time t. We believe that accounting for all the past visit information may help the predictive models to improve the accuracy and provide better interpretation.

To tackle all the aforementioned issues and challenges, we propose an efficient and accurate **di**agnosis **p**rediction m**o**d**el** (Dipole) using attention-based bidirectional recurrent neural networks for learning low-dimensional representations of the patient visits, and employ the learned representations for future diagnosis prediction. The learned representations are easily interpretable, and can also be used to learn how important each visit is to the future diagnosis prediction. Specifically, it first embeds the high dimensional medical codes (i.e., clinical variables) into a low code-level space, and then feeds the code representations into an attention-based bidirectional recurrent neural network to generate the hidden state representation. The hidden representation is fed through a softmax layer to predict the medical codes in future visits. We experiment with three types of attention mechansims: (i) location-based, (ii) general, and (iii) concatenation-based, to calculate the attention weights for all the prior visits for each patient. These mechanisms model the inter-visit relationships, where the attention weights represent the importance of each visit.

We demonstrate that the proposed Dipole achieves significantly higher prediction accuracy when compared to the state-of-the-art approaches in diagnosis prediction, using two datasets derived from Medicaid claims data. A case study is conducted to show that the proposed model accurately assigns varying attention weights to past visits. We evaluate the interpretability of the learned representations through qualitative analysis. Finally, we illustrate the reasonableness of employing bidirectional recurrent neural networks to model temporal patient visits. In summary, our main contributions are as follows:

- We propose Dipole, an end-to-end, simple and robust model to accurately predict the future visit information and reasonably interpret the prediction results, without depending on any expert medical knowledge.
- Dipole models patient visit information in a time-ordered and reverse time-ordered way and employs three attention mechanisms to calculate the weights for previous visits.
- We empirically show that the proposed Dipole outperforms existing methods in diagnosis prediction on two large real world EHR datasets.
- We analyze the experimental results with clinical experts to validate the interpretability of the learned medical code representations.

The rest of this paper is organized as follows: In Section 2, we discuss the connection of the proposed approaches to related work. Section 3 presents the details of the proposed Dipole. The experimental results are presented in Section 4. Section 5 concludes the paper.

2 RELATED WORK

This section reviews the existing work for mining Electronic Healthcare Records (EHR) data. In particular, we focus on the state-of-the-art models on diagnosis prediction task. We also introduce some work using attention mechanisms.

2.1 EHR Data Mining

Mining EHR data is a hot research topic in healthcare informatics. The tackled tasks include electronic genotyping and phenotyping [5, 16, 19, 32], disease progression [12, 26, 27, 33], adverse drug event detection [21], diagnosis prediction [9–11, 25, 31], and so on. In most tasks, deep learning models can significantly improve the performance. Recurrent neural networks (RNNs) can be used for modeling multivariate time series data in healthcare with missing values [6, 18]. Convolutional neural networks (CNNs) are used to predict unplanned readmission [23] and risk [7] with EHR. Stacked denoising autoencoders (SDAs) are employed to detect the characteristic patterns of physiology in clinical time series data [5].

Diagnosis prediction is an important and difficult task in healthcare. Med2Vec [9] aims to learn the representations of medical codes, which can be used to predict the future visit information. This method ignores long-term dependencies of medical codes among visits. RETAIN [11] is an interpretable predictive model, which employs reverse time attention mechanism in an RNN for binary prediction task. GRAM [10] is a graph-based attention model for healthcare representation learning, which uses medical ontologies to learn robust representations and an RNN to model patient visits. Both RETAIN and GRAM apply attention mechanisms and improve the prediction performance.

Compared with the aforementioned predictive models, the proposed approaches not only employ bidirectional neural networks when modeling visit information but also design different attention mechanisms to assign different weights for the past visits. Relying on these two properties, the proposed Dipole can improve the prediction performance significantly and interpret the meanings of medical codes reasonably.

2.2 Attention-based Neural Networks

Attention-based neural networks have been successfully used in many tasks [1, 2, 13, 14, 17, 20, 28, 29]. Specifically for neural machine translation [2, 20], given a sentence in the original language

(i.e., original space), RNNs were adopted to generate the word representations in the sentence $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_{|S|}$, where $|S|$ is the number of words in this sentence. In order to find the t-th word in the target language (or target space), a weight α_{ti}, i.e., attention score, is assigned to each word in the original language. Then, a context vector $\boldsymbol{c}_t = \sum_{i=1}^{|S|} \alpha_{ti} \boldsymbol{h}_i$ is calculated to predict the t-th word in the target language. However, diagnosis prediction task is different from the language translation task as all the visits for each patient are in the same space.

3 METHODOLOGY

In this section, we first introduce the structure of EHR data and some basic notations. Then we describe the details of the proposed Dipole neural network. Finally, we describe the interpretation for the learned code representations and visit representations.

3.1 Basic Notations

We denote all the unique medical codes from the EHR data as $c_1, c_2, \cdots, c_{|C|} \in C$, where $|C|$ is the number of unique medical codes. Assuming there are N patients, the n-th patient has $T^{(n)}$ visit records in the EHR data. The patient can be represented by a sequence of visits $V_1, V_2, \cdots, V_{T^{(n)}}$. Each visit V_t, containing a subset of medical codes ($V_t \subseteq C$), is denoted by a binary vector $\boldsymbol{x}_t \in \{0,1\}^{|C|}$, where the i-th element is 1 if V_t contains the code c_i. Each diagnosis code can be mapped to a node of the International Classification of Diseases (ICD-9)[1], and each procedure code can be mapped to a node in the Current Procedural Terminology (CPT)[2]. Below we use a simple example to illustrate the problem.

There are two diagnosis codes: 250 (*Diabetes mellitus*) and 254 (*Diseases of thymus gland*), and one procedure code 11720 (*Debride nail, 1-5*) in the whole dataset, i.e., $|C| = 3$. If the medical codes in the t-th patient visit are 250 and 254, then $\boldsymbol{x}_t = [1, 1, 0]$.

Both ICD-9 and CPT systems are coded hierarchically, which means that each medical code has a "parent", i.e., category label. For example, the diagnosis codes 250 and 254 belong to the same category *Diseases of other endocrine glands*, and the procedure code 11720 is in the category *Surgical procedures on the nails*. Thus, each visit V_t has a corresponding coarse-grained category representation $\boldsymbol{y}_t \in \{0,1\}^{|\mathcal{G}|}$, where $|\mathcal{G}|$ is the unique number of categories. In the above example, $|\mathcal{G}| = 2$, and $\boldsymbol{y}_t = [1, 0]$. For simplicity, we describe the proposed algorithm for a single patient and drop the superscript (n) when it is unambiguous. The input of the proposed Dipole model is a time-ordered sequence of patient visits.

3.2 Model

The goal of the proposed algorithm is to predict the $(t+1)$-th visit's category-level medical codes. Figure 1 shows the high-level overview of the proposed model. Given the visit information from time 1 to t, the i-th visit's medical codes $\boldsymbol{x}_i$ can be embedded into a vector representation $\boldsymbol{v}_i$. The vector $\boldsymbol{v}_i$ is fed into the Bidirectional Recurrent Neural Network (BRNN) [24], which outputs a hidden state $\boldsymbol{h}_i$ as the representation of the i-th visit. Along with the set of hidden states $\{\boldsymbol{h}_i\}_{i=1}^{t-1}$, we are able to compute the relative importance vector $\boldsymbol{\alpha}_t$ for the current visit t. Subsequently, a *context* state $\boldsymbol{c}_t$ is computed from the relative importance $\boldsymbol{\alpha}_t$ and $\{\boldsymbol{h}_i\}_{i=1}^{t-1}$. This procedure is known as attention model [2], which will be detailed in the following sections. Next, from the context state $\boldsymbol{c}_t$ and the current hidden state $\boldsymbol{h}_t$, we can obtain an *attentional hidden* state $\tilde{\boldsymbol{h}}_t$, which is used to predict the category-level medical codes appearing in the $(t+1)$-th visit, i.e., $\hat{\boldsymbol{y}}_t$. The proposed neural network can be trained end-to-end.

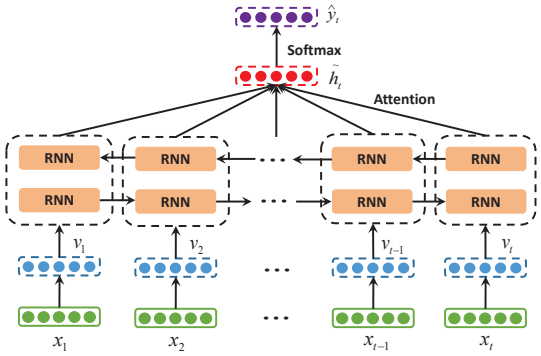


Figure 1: The Proposed Dipole Model.

Visit Embedding

Given a visit $\boldsymbol{x}_i \in \{0,1\}^{|C|}$, we can obtain its vector representation $\boldsymbol{v}_i \in \mathbb{R}^m$ as follows:

$$\boldsymbol{v}_i = \text{ReLU}(\boldsymbol{W}_v \boldsymbol{x}_i + \boldsymbol{b}_c), \tag{1}$$

where m is the size of embedding dimension, $\boldsymbol{W}_v \in \mathbb{R}^{m \times |C|}$ is the weight matrix of medical codes, and $\boldsymbol{b}_c \in \mathbb{R}^m$ is the bias vector. ReLU is the rectified linear unit defined as $\text{ReLU}(\boldsymbol{v}) = \max(\boldsymbol{v}, \boldsymbol{0})$, where max() applies element-wise to vectors. The reason we employ the rectified linear unit as the activation function is that ReLU enables the learned vector representations to be interpretable [9].

Bidirectional Recurrent Neural Networks

Recurrent Neural Networks (RNNs) provide a very elegant way of modeling sequential healthcare data [10, 11]. However, one drawback of RNNs is that the prediction performance will drop when the length of the sequence is very large [24]. In order to overcome this drawback, we employ Bidirectional Recurrent Neural Networks (BRNNs) in the proposed model which can be trained using all the available input visits' information from two directions to improve the prediction performance. Note that we use "RNNs" to denote any Recurrent Neural Networks variant dealing with the vanishing gradient problem [3], such as Long-Short Term Memory (LSTM) [15] and Gated Recurrent Unit (GRU) [8]. In our implementation, we use GRU to adaptively capture dependencies among patient visit information.

A BRNN consists of a forward and backward RNN. The forward RNN $\overrightarrow{f}$ reads the input visit sequence from $\boldsymbol{x}_1$ to $\boldsymbol{x}_T$ and calculates a sequence of *forward hidden states* $(\overrightarrow{\boldsymbol{h}}_1, \cdots, \overrightarrow{\boldsymbol{h}}_T)$ $(\overrightarrow{\boldsymbol{h}}_i \in \mathbb{R}^p$ and p is the dimensionality of hidden states). The backward RNN $\overleftarrow{f}$ reads the visit sequence in the reverse order, i.e., from $\boldsymbol{x}_T$ to $\boldsymbol{x}_1$, resulting in a sequence of *backward hidden states* $(\overleftarrow{\boldsymbol{h}}_1, \cdots, \overleftarrow{\boldsymbol{h}}_T)$ $(\overleftarrow{\boldsymbol{h}}_i \in \mathbb{R}^p)$.

[1] https://en.wikipedia.org/wiki/List_of_ICD-9_codes
[2] https://en.wikipedia.org/wiki/Current_Procedural_Terminology

By concatenating the forward hidden state $\overrightarrow{\boldsymbol{h}}_i$ and the backward one $\overleftarrow{\boldsymbol{h}}_i$, we can obtain the final latent vector representation as $\boldsymbol{h}_i = [\overrightarrow{\boldsymbol{h}}_i; \overleftarrow{\boldsymbol{h}}_i]^\top$ ($\boldsymbol{h}_i \in \mathbb{R}^{2p}$). Note that the future visit information is only used when training the model. Only the past visit information is provided to predict the future visits during the testing phase.

Attention Mechanism

In diagnosis prediction task, the final goal is to predict the category-level medical codes of the $(t+1)$-th visit, i.e., $\boldsymbol{y}_t$, according to the visits from $\boldsymbol{x}_1$ to $\boldsymbol{x}_t$. The output of the t-th visit $\boldsymbol{x}_t$ ($\boldsymbol{h}_t$) is the estimated vector representation of the $(t+1)$-th visit. However, it may contain partial visit information to be predicted. Thus, how to derive a context vector $\boldsymbol{c}_t$ that captures relevant information to help predict the future visit $\boldsymbol{y}_t$ is the key issue. There are three methods that can be used to compute the context vector $\boldsymbol{c}_t$:

• *Location-based Attention.* A location-based attention function is to calculate the weights solely from the current hidden state $\boldsymbol{h}_i$ as follows:

$$\alpha_{ti} = \boldsymbol{W}_\alpha^\top \boldsymbol{h}_i + b_\alpha, \tag{2}$$

where $\boldsymbol{W}_\alpha \in \mathbb{R}^{2p}$ and $b_\alpha \in \mathbb{R}$ are the parameters to be learned. According to Eq. (2), we can obtain an attention weight vector $\boldsymbol{\alpha}_t$ using softmax function as follows:

$$\boldsymbol{\alpha}_t = \text{Softmax}([\alpha_{t1}, \alpha_{t2}, \cdots, \alpha_{t(t-1)}]). \tag{3}$$

Then the context vector $\boldsymbol{c}_t \in \mathbb{R}^{2p}$ can be calculated based on the weights obtained from Eq. (3) and the hidden states from $\boldsymbol{h}_1$ to $\boldsymbol{h}_{t-1}$ as follows:

$$\boldsymbol{c}_t = \sum_{i=1}^{t-1} \alpha_{ti} \boldsymbol{h}_i. \tag{4}$$

Since location-based attention mechanism only considers each individual hidden state information, it does not capture the relationships between the current hidden state and all the previous hidden states. To utilize the information from all the previous hidden states, we adopt the following two attention mechanisms in the proposed Dipole .

• *General Attention.* An easy way to capture the relationship between $\boldsymbol{h}_t$ and $\boldsymbol{h}_i$ ($1 \leqslant i \leqslant t-1$) is using a matrix $\boldsymbol{W}_\alpha \in \mathbb{R}^{2p \times 2p}$, and calculating the weight as:

$$\alpha_{ti} = \boldsymbol{h}_t^\top \boldsymbol{W}_\alpha \boldsymbol{h}_i, \tag{5}$$

and the context vector $\boldsymbol{c}_t$ can be obtained using Eq. (3) and Eq. (4).

• *Concatenation-based Attention.* Another way to calculate the context vector $\boldsymbol{c}_t$ is using a multi-layer perceptron (MLP) [2]. We first concatenate the current hidden state $\boldsymbol{h}_s$ and the previous state $\boldsymbol{h}_i$, and then a latent vector can be obtained by multiplying a weight matrix $\boldsymbol{W}_\alpha \in \mathbb{R}^{q \times 4p}$, where q is the latent dimensionality. We select *tanh* as the activation function. The attention weight vector is generated as follows:

$$\alpha_{ti} = \boldsymbol{v}_\alpha^\top \tanh(\boldsymbol{W}_\alpha [\boldsymbol{h}_t; \boldsymbol{h}_i]), \tag{6}$$

where $\boldsymbol{v}_\alpha \in \mathbb{R}^q$ is the parameter to be learned, and we can obtain the context vector $\boldsymbol{c}_t$ with Eq. (3) and Eq. (4).

Diagnosis Prediction

Given the context vector $\boldsymbol{c}_t$ and the current hidden state $\boldsymbol{h}_t$, we employ a simple concatenation layer to combine the information from both vectors to generate an attentional hidden state as follows:

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_c [\boldsymbol{c}_t; \boldsymbol{h}_t]), \tag{7}$$

where $\boldsymbol{W}_c \in \mathbb{R}^{r \times 4p}$ is the weight matrix. The attentional vector $\tilde{\boldsymbol{h}}_t$ is fed through the softmax layer to produce the $(t+1)$-th visit information defined as:

$$\hat{\boldsymbol{y}}_t = \text{Softmax}(\boldsymbol{W}_s \tilde{\boldsymbol{h}}_t + \boldsymbol{b}_s) \tag{8}$$

where $\boldsymbol{W}_s \in \mathbb{R}^{|\mathcal{G}| \times r}$ and $\boldsymbol{b}_s \in \mathbb{R}^{|\mathcal{G}|}$ are the parameters to be learned.

Objective Function

Based on Eq. (8), we use the cross-entropy between the ground truth visit information $\boldsymbol{y}_t$ and the predicted visit $\hat{\boldsymbol{y}}_t$ to calculate the loss for all the patients as follows:

$$\begin{aligned} &\mathcal{L}(\boldsymbol{x}_1^{(1)}, \cdots, \boldsymbol{x}_{T^{(1)}-1}^{(1)}, \cdots, \boldsymbol{x}_1^{(N)}, \cdots, \boldsymbol{x}_{T^{(N)}-1}^{(N)}) \\ &= -\frac{1}{N} \sum_{n=1}^{N} \frac{1}{T^{(n)}-1} \sum_{t=1}^{T^{(n)}-1} \left(\boldsymbol{y}_t^\top \log(\hat{\boldsymbol{y}}_t) + (\boldsymbol{1} - \boldsymbol{y}_t)^\top \log(\boldsymbol{1} - \hat{\boldsymbol{y}}_t) \right) \end{aligned} \tag{9}$$

3.3 Interpretation

In healthcare, the interpretability of the learned representations of medical codes and visits is important. We need to understand the clinical meaning of each dimension of medical code representations, and analyze which visits are crucial to the prediction.

Since the proposed model is based on attention mechanisms, it is easy to find the importance of each visit for prediction by analyzing the attention scores. For the t-th prediction, if the attention score α_{ti} is large, then the probability of the $(i+1)$-th visit information related to the current prediction is high. We employ the simple method proposed in [9] to interpret the code representations. We first use $\text{ReLU}(\boldsymbol{W}_v^\top)$, a non-negative matrix, to represent the medical codes. Then we rank the codes by values in a reverse order for each dimension of the hidden state vector. Finally, the top k codes with the largest values are selected as follows:

$$\text{argsort}(\boldsymbol{W}_v^\top[:, i])[1:k],$$

where $\boldsymbol{W}_v^\top[:, i]$ represents the i-th column or dimension of $\boldsymbol{W}_v^\top$. By analyzing the selected medical codes, we can obtain the clinical interpretation of each dimension. Detailed examples and analysis are given in Section 4.4 and 4.5.

4 EXPERIMENTS

In this section, we evaluate the performance of the proposed Dipole model on two real world insurance claims datasets, compare its performance with other state-of-the-art prediction models, and show that it yields higher accuracy.

4.1 Data Description

The datasets we used in the experiments are the Medicaid claims and the Diabetes claims.

The Medicaid Dataset

Our first dataset consists of Medicaid claims[3] over the year 2011. It consists of data corresponding to 147, 810 patients, and 1, 055, 011

[3] https://www.medicaid.gov

visits. The patient visits were grouped by week, and we excluded patients who made less than two visits.

The Diabetes Dataset

The Diabetes dataset consists of Medicaid claims over the years 2012 and 2013, corresponding to patients who have been diagnosed with diabetes (i.e. Medicaid members who have the ICD-9 diagnosis code 250.xx in their claims). It contains data corresponding to 22,820 patients with 466,732 visits. The patient visits were aggregated by week, and excluded patients who made less than five visits.

For both datasets, each visit information includes the ICD-9 diagnosis codes and procedure codes, categorized in accordance with the Current Procedural Terminology (CPT). Table 1 lists more details about the two datasets.

Table 1: Statistics of Diabetes and Medicaid Dataset.

Dataset	Diabetes	Medicaid
# of patients	22,820	147,810
# of visits	466,732	1,055,011
Avg. # of visits per patient	20.45	7.14
# of unique medical codes	7,399	8,522
- # of unique diagnosis codes	984	1,021
- # of unique procedure codes	6,415	7,501
Avg. # of medical codes per visit	6.35	5.57
Max # of medical codes per visit	105	99
# of category codes	422	426
- # of unique diagnosis categories	183	185
- # of unique procedure categories	239	241
Avg. # of category codes per visit	4.85	4.08
Max # of category codes per visit	38	42

4.2 Experimental Setup

In this subsection, we first describe the state-of-the-art approaches for EHR representation learning and diagnosis prediction which are used as baselines, and then outline the measures used for evaluation. Finally, we introduce the implementation details.

Baseline Approaches

To validate the performance of the proposed model for diagnosis prediction task, we compare it with several state-of-the-art models. We select three existing approaches as baselines[4]:

Med2Vec [9]. Med2Vec, which follows the idea of Skip-gram [22], is a simple and robust algorithm to efficiently learn medical code representations and predict the medical codes appearing in the following visit based on the current visit information.

RETAIN [11]. RETAIN is an interpretable predictive model in healthcare with reverse time attention mechanism, a two-level neural attention model. It can find influential past visits and important medical codes within those visits. Since the original RETAIN is used for binary prediction task, we change the final softmax function for satisfying multiple variable prediction, i.e., diagnosis prediction.

RNN. We first embed visit information into vector representations according to Eq. (1), then feed this embedding to the GRU. The hidden states produced by the GRU are directly used to predict the medical codes of the $(t+1)$-th visit using softmax according to Eq. (8). All the parameters are trained together with the GRU.

Our Approaches

Since all the attention mechanisms proposed in Section 3.2 can be used for RNN model, we propose three variants of RNN as follows:

RNN_l. We add location-based attention model into RNN. The attention scores are calculated by Eq. (2). Then we can obtain the context vectors according to Eq. (4). Based on the context vectors, we can generate attention hidden states using Eq. (7). Finally, we can predict the medical codes of the $(t+1)$ visit using Eq. (8).

RNN_g. RNN_g is similar to RNN_l, but uses general attention model, i.e., Eq. (5), to calculate attention scores.

RNN_c. RNN_c uses concatenation-based attention mechanism (Eq. (6)) to calculate attention weights.

The proposed Dipole model is a general framework for predicting diagnoses in healthcare. We show the performance of the following four approaches in the experiments.

Dipole−. This model only uses the hidden states generated by BRNN to predict the next visit information, i.e., without employing any attention mechanisms.

Dipole_l. It is based on location-based attention mechanism with Eq. (2).

Dipole_g. Dipole_g uses general attention model when calculating the context vectors, i.e., Eq. (5).

Dipole_c. Similar to Dipole_l and Dipole_g, Dipole_c employs concatenation based attention mechanism (Eq. (6)) in the predictive model.

Evaluation Strategies

To evaluate the performance of predicting future medical codes for each method, we use two measures: accuracy and accuracy@k . Accuracy is defined as the correct medical codes ranked in top k divided by $|\boldsymbol{y}_t|$, where $|\boldsymbol{y}_t|$ is the number of medical codes in the $(t+1)$-th visit, and k equals to $|\boldsymbol{y}_t|$. Accuracy@k is defined as the correct medical codes in top k divided by $\min(k, |\boldsymbol{y}_t|)$. In our experiments, we vary k from 5 to 30.

Implementation Details

We implement all the approaches with Theano 0.7.0 [4]. For training models, we use Adadelta [30] with a mini-batch of 100 patients[5]. We randomly divide the dataset into the training, validation and testing set in a 0.75:0.1:0.15 ratio. The validation set is used to determine the best values of parameters. We also use regularization (l_2 norm with the coefficient 0.001) and drop-out strategies (the drop-out rate is 0.5) for all the approaches. In the experiments, we set the same $m = 256$, $p = 256$ and $q = 128$ for baselines and our approaches. We perform 100 iterations and report the best performance for each method.

4.3 Results of Diagnosis Prediction

Table 2 shows the accuracy of the proposed approaches and baselines on both Diabetes and Medicaid datasets for the diagnosis

[4]GRAM [10] is not a baseline as it uses external knowledge to learn the medical code representations.

[5]For Med2Vec, we use 1000 visits per batch as in [9]. The window size is 5 on the Diabetes dataset and 3 on the Medicaid dataset.

prediction task. #C represents the average number of correct predictions. The number of visits or predictions in the test set is 65,975 in the Diabetes dataset, and 136,023 in the Medicaid dataset.

Table 2: The Accuracy of Diagnosis Prediction Task.

Method		Diabetes		Medicaid	
		# C	Accuracy	# C	Accuracy
Baseline	RNN	28,608	0.4336	58,245	0.4282
	Med2Vec	29,175	0.4422	62,326	0.4582
	RETAIN	28,851	0.4373	63,496	0.4668
Our Approach	RNN_l	29,880	0.4529	62,938	0.4627
	RNN_g	30,124	0.4566	62,571	0.4600
	RNN_c	30,164	0.4572	62,693	0.4609
	Dipole−	29,623	0.4490	62,475	0.4593
	Dipole_l	30,645	0.4645	**65,441**	**0.4811**
	Dipole_g	29,464	0.4466	64,557	0.4746
	Dipole_c	**30,698**	**0.4653**	63,931	0.4700

In Table 2, we can observe that the accuracy of the proposed approaches, including Dipole and RNN variants, is higher than that of baselines on the Diabetes dataset. Since most medical codes are about diabetes, Med2Vec can correctly learn vector representations on the Diabetes dataset. Thus, Med2Vec achieves the best results among the three baselines. For the Medicaid dataset, the accuracy of RETAIN is better than that of Med2Vec. The reason is that there are many diseases in the Medicaid dataset, and the categories of medical codes are more than those on the Diabetes dataset. In this case, attention mechanism can help RETAIN to learn reasonable parameters and make correct prediction.

The accuracy of RNN is the lowest among all the approaches on both datasets. This is because the prediction of RNN mostly depends on the recent visits' information. It cannot memorize all the past information. However, RETAIN and the proposed RNN variants, RNN_l, RNN_g and RNN_c, can fully take all the previous visit information into consideration, assign different attention scores for past visits, and achieve better performance when compared to RNN.

Since most of the visits on the Diabetes dataset are related to diabetes, it is easy to predict the medical codes in the next visit according to the past visit information. RETAIN uses a reverse time attention mechanism for prediction, which will decrease the prediction performance compared with the approaches using a time ordered attention mechanism. Thus, the performance of the three proposed RNN variants is better than that of RETAIN. However, the accuracy of RETAIN is better than the proposed RNN variants' as the data are about different diseases on the Medicaid dataset. Using the reverse time attention mechanism can help us to learn the correct relationships among visits.

Both RNN and the proposed Dipole− do not use any attention mechanism, but the accuracy of Dipole− is higher than that of RNN on both the Diabetes and Medicaid dataset. It shows that modeling visit information from two directions can improve the prediction performance. Thus, it is reasonable to employ bidirectional recurrent neural networks for diagnosis prediction task.

The proposed Dipole_c and Dipole_l can achieve the best performance on the Diabetes and Medicaid dataset respectively, which shows that both modeling visits from two directions and assigning a different weight to each visit can improve the accuracy for diagnosis prediction task in healthcare. On the Diabetes dataset, Dipole_l and Dipole_c outperform all the baselines and the proposed RNN variants. On the Medicaid dataset, the performance of all the three proposed approaches, Dipole_l, Dipole_g and Dipole_c is better than that of baselines and RNN variants.

Table 3 shows the experimental results with the accuracy@k measurement on the Diabetes and Medicaid dataset separately. We can observe that as k increases, the performance of all the approaches improves, but the proposed Dipole approaches show superior predictive performance, demonstrating their applicability in predictive healthcare modeling. In Table 3, RETAIN can achieve comparable performance with the proposed approaches on the Medicaid dataset. The overall performance of location-based attention methods, Dipole_l and RNN_l, is better than that of other methods, which indicates that location-based attention performs well on this dataset. RETAIN also uses location-based attention mechanism. Thus, it can obtain high accuracy.

4.4 Case Study

To demonstrate the benefit of applying attention mechanisms in diagnosis prediction task, we analyze the attention weights learned from one of the proposed approach Dipole_c, which uses concatenation based attention mechanism. Figure 2 shows a case study for predicting the medical code in the sixth visit ($\boldsymbol{y}_5$) based on the previous visits on the Diabetes dataset. The concatenation-based attention weights are calculated for the visits from the second visit to the fifth visit according to the hidden states $\boldsymbol{h}_1$, $\boldsymbol{h}_2$, $\boldsymbol{h}_3$ and $\boldsymbol{h}_4$. Thus, we have four attention scores. In Figure 2, X-axis represents patients, and Y-axis is the attention weight calculated for each visit. In this case study, we select five patients. We can observe that for different patients, the attention scores learned by the attention mechanism are different.

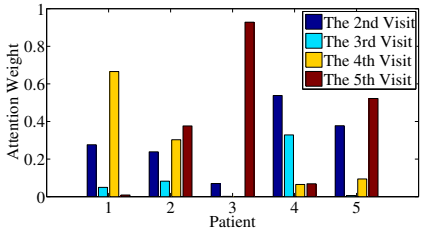


Figure 2: Attention Mechanism Analysis.

To illustrate the correctness of the learned attention weights, we provide an example. For the second patient in Figure 2, we list all the diagnosis codes in Table 4. In order to predict the medical codes in the sixth visits, we first obtain the attention scores $\boldsymbol{\alpha} = [0.2386, 0.0824, 0.3028, 0.3762]$. Analyzing this attention vector, we can conclude that the medical codes in the second, fourth and fifth visits significantly contribute to the final prediction. From Table 4, we can observe that the patient suffered *essential hypertension* in the second and fourth visits, and diagnosed *diabetes* in the fifth visits. Thus, the probability of the sixth visit's medical codes about *diabetes* and diseases related to *essential hypertension* is high.

Table 3: The Accuracy@k of Diagnosis Prediction Task.

Dataset	Accuracy@k	RNN	Med2Vec	RETAIN	RNN$_l$	RNN$_g$	RNN$_c$	Dipole−	Dipole$_l$	Dipole$_g$	Dipole$_c$
Diabetes	5	0.5236	0.5210	0.5257	0.5389	0.5423	0.5418	0.5413	0.5568	0.5350	**0.5575**
	10	0.6107	0.6015	0.6102	0.6281	0.6316	0.6317	0.6325	0.6466	0.6204	**0.6469**
	15	0.6829	0.6744	0.6826	0.6993	0.7026	0.7034	0.7036	0.7146	0.6913	**0.7150**
	20	0.7355	0.7257	0.7354	0.7515	0.7542	0.7555	0.7548	**0.7642**	0.7415	0.7641
	25	0.7759	0.7662	0.7761	0.7928	0.7949	0.7959	0.7942	**0.8021**	0.7820	0.8019
	30	0.8087	0.7998	0.8091	0.8259	0.8273	0.8279	0.8251	**0.8318**	0.8141	0.8316
Medicaid	5	0.5238	0.5470	0.5663	0.5579	0.5540	0.5569	0.5586	**0.5791**	0.5698	0.5660
	10	0.6237	0.6342	0.6620	0.6645	0.6595	0.6627	0.6575	**0.6764**	0.6663	0.6615
	15	0.6933	0.7015	0.7297	0.7348	0.7300	0.7329	0.7249	**0.7420**	0.7324	0.7268
	20	0.7444	0.7511	0.7773	0.7842	0.7809	0.7828	0.7720	**0.7877**	0.7785	0.7736
	25	0.7843	0.7902	0.8134	0.8211	0.8183	0.8197	0.8074	**0.8213**	0.8127	0.8088
	30	0.8157	0.8211	0.8416	**0.8496**	0.8469	0.8482	0.8358	0.8475	0.8400	0.8362

Table 4: Diagnosis Codes in Each Visit for Patient 2 in the Case Study.

Visit	Diagnosis Codes
1	Symptoms involving digestive system (787) Essential hypertension (401) Symptoms involving respiratory system and other chest symptoms (786) Special screening for other conditions (V82)
2	Essential hypertension (401)
3	Disorders of lipoid metabolism (272) Hypotension (458)
4	Essential hypertension (401) Need for isolation and other prophylactic measures (V07)
5	Diabetes mellitus (250)
6	Hypertensive heart disease (402) Diabetes mellitus (250)

According to the proposed approach, we can predict the correct diagnoses that this patient suffers *diabetes* and *hypertensive heart disease*. This case study demonstrates that we can learn an accurate attention weight for each visit, and the experimental results in Section 4.3 also illustrate that the appropriate attention models can significantly improve the performance of the diagnosis prediction task in healthcare.

4.5 Code Representation Analysis

The interpretability of medical codes is important in healthcare. In order to analyze the representations of medical codes learned by the proposed model Dipole$_g$, we show top ten diagnosis codes with the largest value in each of six columns selected from the hidden representation matrix $\boldsymbol{W}_v^\top \in \mathbb{R}^{|C|\times m}$ in Table 5. In this way, we can demonstrate the characteristic of each column and map each dimension from the code embedding space to the medical concept.

In Table 5, we can clearly observe that the codes in all the six dimensions are about diabetes complications, which are in accordance with the complications listed on the American Diabetes Association[6]. Dimension 10 is related to eye complications and Alzheimer's disease. Diabetes can damage the blood vessels of the retina (diabetic retinopathy), potentially leading to blindness, and Type 2 diabetes may increase the risk of Alzheimer's disease. Dimension 38 relates to the complications of neuropathy (nerve damage). Dimension 77 is about heart diseases. It has been established that there is a high correlation between diabetes, heart disease, and stroke. In fact, two out of three patients with diabetes die from heart disease or stroke. Patients with diabetes have a greater risk of depression than people without diabetes. Dimension 79 includes the codes related to mental health. Dimension 141 shows a fact that diabetes may cause skin problems, including bacterial and fungal infections. High blood pressure is one common complication of diabetes shown in dimension 142, which also raises the risk for heart attack, stroke, eye problems, and kidney disease.

[6]http://www.diabetes.org/living-with-diabetes/complications/

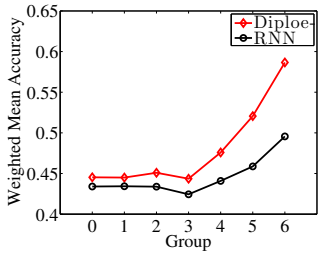
4.6 Assumption Validation

In the proposed model, we adopt bidirectional recurrent neural networks to model patient visits instead of recurrent neural networks. To illustrate the benefit of employing bidirectional recurrent neural networks, we analyze the detailed mean accuracy of RNN and Dipole- shown in Figure 3. We first divide patients into different groups based on the number of visits. The group label is the quotient of the number of visits divided by 15 for the Diabetes dataset and 7 for the Medicaid dataset, which is X-axis in Figure 3. Then we calculate the weighted average accuracy (Y-axis) of different groups, i.e., $\frac{\sum_n MA_n * C_n}{\sum_n C_n}$, where MA_n is the mean accuracy of all the patients with n visits, and C_n is the number of patients with n visits. From Figure 3, we can observe that the average accuracy of Dipole- is better than that of RNN in different groups. On the Diabetes dataset, the weighted mean accuracy of RNN increases when the number of visits becomes larger. This is because the codes of visits on the Diabetes dataset are all about diabetes, and RNN can make correct prediction according to recent visits' information. However, the codes on the Medicaid dataset are related to multiple diseases, and it is hard to correctly predict the future visit information when the sequences are too long. Thus, the weighted mean accuracy significantly drops when the number of visits is large on the Medicaid dataset.

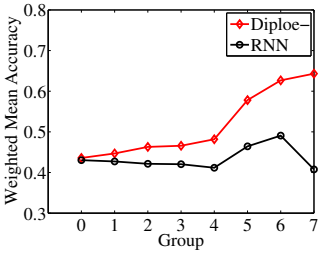
Figure 4 shows the difference of weighted mean accuracy between Dipole- and RNN in different groups. We can observe that

Table 5: Interpretation for Diagnosis Code Representations on the Diabetes Dataset.

Dimension 10	Dimension 38	Dimension 77
Glaucoma (365) Fracture of one or more tarsal and metatarsal bones (825) Dementias (290) Psoriasis and similar disorders (696) Mild mental retardation (317) Cataract (366) Injury, other and unspecified (959) Rheumatoid arthritis and other inflammatory polyarthropathies (714) Thyrotoxicosis with or without goiter(242) Blindness and low vision (369)	Hereditary and idiopathic peripheral neuropathy (356) Other disorders of soft tissues (729) Dermatophytosis (110) Other disorders of urethra and urinary track (599) Mononeuritis of lower limb (355) Diabetes mellitus (250) Mononeuritis of upper limb and mononeuritis multiplex (354) Sprains and strains of sacroiliac region (846) Osteoarthrosis and allied disorders (715) Other and unspecified disorders of back (724)	Cardiac dysrhythmias (427) Chronic pulmonary heart disease (416) Special screening for malignant neoplasms (V76) Angina pectoris (413) Other hernia of abdominal cavity without mention of obstruction (553) Cardiomyopathy (425) Ill-defined descriptions and complications of heart disease (429) Diabetes mellitus (250) Acute pulmonary heart disease (415) Gastrointestinal hemorrhage (578)
Dimension 79	**Dimension 141**	**Dimension 142**
Neurotic disorders (300) Other current conditions in the mother classifiable elsewhere (648) Symptoms concerning nutrition metabolism and development (783) Obesity and other hyperalimentation (278) Diseases of esophagus (530) Other organic psychotic conditions (chronic) (294) Schizophrenic disorders (295) Asthma (493) Chronic liver disease and cirrhosis (571) Spondylosis and allied disorders (721)	Viral hepatitis (070) Other cellulitis and abscess (682) Other personal history presenting hazards to health (V15) Cellulitis and abscess of finger and toe (681) Bacterial infection in conditions classified elsewhere (041) Episodic mood disorders (296) Chronic ulcer of skin (707) Mononeuritis of upper limb and mononeuritis multiplex (354) Other diseases due to viruses and Chlamydiae (078) Diabetes mellitus (250)	Essential hypertension (401) Hypertensive renal disease (403) Hypertensive heart disease (402) Chronic renal failure (585) Other disorders of kidney and ureter (593) Other psychosocial circumstances (V62) Secondary hypertension (405) Nonspecific abnormal results of function studies (794) Calculus of kidney and ureter (592) Other organic psychotic conditions (chronic) (294)

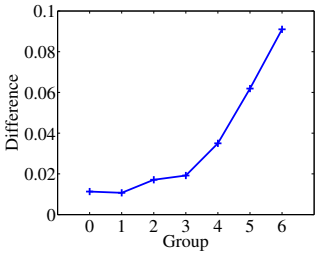


(a) Diabetes Dataset

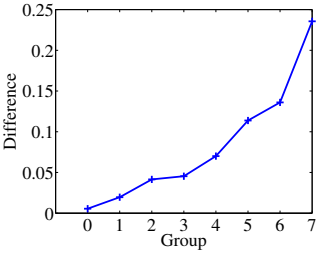


(b) Medicaid Dataset

Figure 3: Weighted Mean Accuracy of Different Groups.



(a) Diabetes Dataset



(b) Medicaid Dataset

Figure 4: Difference of Weighted Mean Accuracy.

with the increase of the number of visits, the difference also augments dramatically. It demonstrates that bidirectional recurrent neural networks can "remember" more information when the sequences are long, and make correct predictions with their memories. Thus, modeling patient visits with bidirectional recurrent neural networks is reasonable.

5 CONCLUSIONS

Diagnosis prediction is a challenging and important task, and interpreting the prediction results is a hard and vital problem for predictive model in healthcare. Many existing work in diagnosis prediction employs deep learning techniques, such as recurrent neural networks (RNNs), to model the temporal and high dimensional EHR data. However, RNN-based approaches may not fully

remember all the previous visit information, which leads to the incorrect prediction. To interpret the predicting results, existing work introduces location-based attention model, but this mechanism ignores the relationships between the current visit and the past visits.

In this paper, we propose a novel model, named Dipole, to address the challenges of modeling EHR data and interpreting the prediction results. By employing bidirectional recurrent neural networks, Dipole can remember the hidden knowledge learned from the previous and future visits. Three attention mechanisms allow us to interpret the prediction results reasonably. Experimental results on two large real world EHR datasets prove the effectiveness of the proposed Dipole for diagnosis prediction task. Analysis shows that the attention mechanisms can assign different weights to previous visits when predicting the future visit information. We demonstrate that the learned representations of medical codes are meaningful. Finally, an experiment is conducted to validate the reasonableness and effectiveness of modeling patient visits with bidirectional recurrent neural networks.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions. This work is supported in part by the US National Science Foundation under grants IIS-1319973, IIS-1553411 and IIS-1514204. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. 2015. Multiple Object Recognition with Visual Attention. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*.

[3] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning Long-Term Dependencies with Gradient Descent is Difficult. *IEEE Transactions on Neural Networks* 5, 2 (1994), 157–166.

[4] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: A CPU and GPU Math Compiler in Python. In *Proceedings of the 9th Python in Science Conference (SciPy'10)*. 1–7.

[5] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. 2015. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*. ACM, 507–516.

[6] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2016. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *arXiv preprint arXiv:1606.01865* (2016).

[7] 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 (SDM'16)*. 432–440.

[8] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv preprint arXiv:1409.1259* (2014).

[9] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer Representation Learning for Medical Concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. 1495–1504.

[10] 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 (KDD'17)*. ACM.

[11] 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 (NIPS'16)*. 3504–3512.

[12] Edward Choi, Nan Du, Robert Chen, Le Song, and Jimeng Sun. 2015. Constructing Disease Network and Temporal Progression Model via Context-sensitive Hawkes Process. In *2015 IEEE International Conference on Data Mining (ICDM'15)*. IEEE, 721–726.

[13] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based Models for Speech Recognition. In *Advances in Neural Information Processing Systems (NIPS'15)*. 577–585.

[14] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems (NIPS'15)*. 1693–1701.

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[16] Peter B Jensen, Lars J Jensen, and Søren Brunak. 2012. Mining Electronic Health Records: Towards Better Research Applications and Clinical Care. *Nature Reviews Genetics* 13, 6 (2012), 395–405.

[17] Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor Forcing: A New Algorithm for Training Recurrent Networks. In *Advances In Neural Information Processing Systems (NIPS'16)*. 4601–4609.

[18] Zachary C Lipton, David C Kale, and Randall Wetzel. 2016. Modeling Missing Data in Clinical Time Series with RNNs. In *Proceedings of Machine Learning for Healthcare (MLHC'16)*.

[19] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. 2015. Temporal Phenotyping from Longitudinal Electronic Health Records: A Graph based Framework. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*. ACM, 705–714.

[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*. 1412–1421.

[21] Fenglong Ma, Chuishi Meng, Houping Xiao, Qi Li, Jing Gao, Lu Su, and Aidong Zhang. 2017. Unsupervised Discovery of Drug Side-Effects from Heterogeneous Data Sources. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. ACM.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems (NIPS'13)*. 3111–3119.

[23] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2016. Deepr: A Convolutional Net for Medical Records. *IEEE Journal of Biomedical and Health Informatics* (2016).

[24] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[25] Qiuling Suo, Fenglong Ma, Giovanni Canino, Jing Gao, Aidong Zhang, Pierangelo Veltri, and Agostino Gnasso. 2017. A Multi-task Framework for Monitoring Health Conditions via Attention-based Recurrent Neural Networks. In *Proceedings of the AMIA 2017 Annual Symposium (AMIA'17)*.

[26] Xiang Wang, David Sontag, and Fei Wang. 2014. Unsupervised Learning of Disease Progression Models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. ACM, 85–94.

[27] Houping Xiao, Jing Gao, Long Vu, and Deepak S. Turaga. 2017. Learning Temporal State of Diabetes Patients via Combining Behavioral and Demographic Data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. ACM.

[28] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML'15)*. CoRR, 2048–2057.

[29] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning with Semantic Attention. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 4651–4659.

[30] Matthew D Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *arXiv preprint arXiv:1212.5701* (2012).

[31] Jiayu Zhou, Jimeng Sun, Yashu Liu, Jianying Hu, and Jieping Ye. 2013. Patient Risk Prediction Model via Top-k Stability Selection. In *Proceedings of the 13th SIAM International Conference on Data Mining (SDM'13)*. SIAM, 55–63.

[32] Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. 2014. From Micro to Macro: Data Driven Phenotyping by Densification of Longitudinal Electronic Medical Records. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. ACM, 135–144.

[33] Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. 2011. A Multi-Task Learning Formulation for Predicting Disease Progression. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*. ACM, 814–822.