Sequential Diagnosis Prediction with Transformer and Ontological Representation

Xueping Peng*, Guodong Long*, Tao Shen*, Sen Wang†, Jing Jiang*
* Australian Artificial Intelligence Institute, FEIT, University of Technology Sydney, Australia
† School of Information Technology and Electrical Engineering, The University of Queensland, Australia
Email: {xueping.peng, guodong.long, tao.shen, jing.jiang}@uts.edu.au, sen.wang@uq.edu.au

***Abstract*—Sequential diagnosis prediction on the Electronic Health Record (EHR) has been proven crucial for predictive analytics in the medical domain. EHR data, sequential records of a patient's interactions with healthcare systems, has numerous inherent characteristics of temporality, irregularity and data insufficiency. Some recent works train healthcare predictive models by making use of sequential information in EHR data, but they are vulnerable to irregular, temporal EHR data with the states of admission/discharge from hospital, and insufficient data. To mitigate this, we propose an end-to-end robust transformer-based model called SETOR, which exploits neural ordinary differential equation to handle both irregular intervals between a patient's visits with admitted timestamps and length of stay in each visit, to alleviate the limitation of insufficient data by integrating medical ontology, and to capture the dependencies between the patient's visits by employing multi-layer transformer blocks. Experiments conducted on two real-world healthcare datasets show that, our sequential diagnoses prediction model SETOR not only achieves better predictive results than previous state-of-the-art approaches, irrespective of sufficient or insufficient training data, but also derives more interpretable embeddings of medical codes. The experimental codes are available at the GitHub repository[1].**

***Index Terms*—Electronic Health Record, Transformer, Ontological Representation, EHR, Neural Ordinary Differential Equation**

I. Introduction

With the rapid growth of the utilization of healthcare information systems during the last few decades, huge volumes of electronic health records (EHR) have been accumulated. The patient EHR data typically consists of a sequence of visit records with irregular admitted intervals, and each visit consists of admission and discharge timestamps and a set of clinical events, such as diagnoses, procedures, medications, etc. [1], [2]. Fig. 1 shows an EHR segment of a patient, which is referred to as the patient journey in the paper. Analyzing the EHR data to benefit the care for a large number of patients has been attracting tremendous attentions from both academia and industry. One of the numerous analytical tasks is to predict the future diagnoses [2]–[6] based on a patient's historical EHR data. For example, [3] and [4] employ recurrent neural networks (RNN) to integrate medical ontology for capturing temporal visits and predicting sequential diagnoses. Ref. [2] predict future diagnoses by utilizing co-occurrence statistics and multiple ontological representations via attention mechanism on EHR data. Although the existing methods have achieved promising results, they are still challenged by the following two limitations.

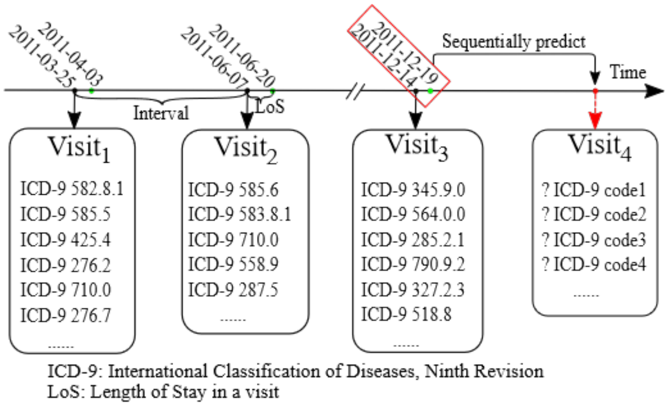


Fig. 1: An EHR segment and prediction task. The black and green dots indicate the admission and discharge dates, respectively.

One of the major challenges is how to effectively model such irregular and temporal EHR data with admission and discharge states from hospital. Some recent works [3]–[5], [7]–[12] directly adapt text representation learning algorithms [13] to the sequence-formatted EHR data. For example, Med2Vec [7] learns a vector representation for each medical concept from the co-occurrence information without considering the temporal sequential nature of the EHR data. Further, considering both long-term dependency and sequential information, recurrent neural networks [3], [4], [10], [11], [14], including LSTM [15] and GRU [16], are used to learn the contextualized representation of EHR data. However, despite the similarity between EHR data and natural language text, one major difference is that EHR data inherently includes the timestamp property. Namely, beyond dependency, there is time interval between each pair of visits. For example, as shown in Fig. 1, the interval between Visit1 and Visit2 is shorter than that between Visit2 and Visit3, indicating that the dependency between Visit1 and Visit2 may be stronger than that later one. Due to irregular visits of patients, this is very common in EMRs. Meanwhile, the previous works, which ignore the discharge states, only consider admitted ones. In other word,

Corresponding: Guodong Long {guodong.long@uts.edu.au}

[1] Github repository: https://github.com/Xueping/SETOR

they rarely take the length of stay in each visit into account for patient journeys.

The other limitation is that existing methods rely on a large volume of training data, which is generally not easily available due to both labour-intensive costs and privacy concerns [17]. Fortunately, some recent works in natural language processing (NLP) field provide effective solutions when task-specific supervised data is scarce. One promising research direction is to leverage off-the-shelter relational knowledge [18], [19] (e.g., Freebase and DBpedia) to enhance the model especially when the knowledge can be used as supportive evidence to the targeted task. Taking this inspiration, recent works [2]–[4] train medical code embeddings upon medical ontology by using graph-based attention mechanism, and thus deliver competitive performance even with insufficient task-specific supervised data. Despite their success in several healthcare tasks, these methods still suffer from a main limitation: rich dependency information underlying the patient's sequential EHR is rarely exploited during medical ontology learning. For example, the medical codes from visit information and medical ontology are heterogeneous, and how to effectively learn both representations and fuse their heterogeneous features is a challenging open task.

To cope with aforementioned limitations, we propose an end-to-end robust transformer-based healthcare analytics model, named **SE**quential Diagnosis Prediction with **T**ransformer and **O**ntological **R**epresentation (SETOR). SETOR integrates medical ontology to alleviate data insufficiency, and exploits neural ordinary differential equation (ODE) [20], [21] to tackle temporal irregularity occurred in both two consecutive visits and length of stay in each visit. Specifically, SETOR first employs the attention-based graph-embedding approach to learn ontological and generalized representations of medical codes to mitigate the problem of data insufficiency. Next, the ontological encoder is proposed to integrate the learned ontological representations into visit information to enhance medical representations. Then, the proposed model utilizes neural ODE to learn the discharge state based on the admitted state for each visit, called LoS (Length of Stay) ODE, and the hidden states for irregular intervals between consecutive patient's visits, called Interval ODE. Lastly, SETOR integrates the learned hidden discharge and interval states and compressed visit vectors to predict sequential diagnoses followed by patient journey transformer. Consequently, the proposed model can improve the prediction quality of future diagnoses, and advance the robustness irrespective of sufficient or insufficient data.

To summarize, our main contributions are:

- novel LoS and interval ODE representations, that use neural ODE to model discharge states and capture the irregular admitted interval dependencies between patient's visits;
- an end-to-end neural network called "SETOR" that accurately predicts sequential diagnoses using neural ODE and ontological representation;
- an evaluation on two real-world datasets, qualitatively demonstrating the interpretability of the learned representations of medical codes and quantitatively validating the effectiveness of the proposed model.

The remainder of this paper is organized as follows. Section II reviews related works. Then, details about our model are presented in Section III. And next, in Section IV, we demonstrate the experimental results conducted on real-world datasets. Lastly, we conclude our work in Section V.

II. Related Work

Deep neural networks have been applied to healthcare analytical tasks, which have recently attracted enormous interest in healthcare informatics. This section reviews two types of related studies, which are sequential prediction and medical ontologies on EHR.

A. *Sequential prediction on EHR*

Sequential prediction of clinical events based on EHR data has been attracting tremendous attentions. Most existing models utilize RNNs and attention mechanism for predicting the future diagnoses. Med2Vec [7] and MIME [22] indirectly exploit an RNN to embed the visit sequences into a patient representation by multi-level representation learning to integrate visits and medical codes based on visit sequences and the co-occurrence of medical codes to predict future health outcomes. Other research works have, however, used RNNs directly to model time-ordered patient visits for predicting diagnoses [3], [4], [10], [11], [14], [23]–[26]. For example, Dipole [10] and RETAIN [11] employ RNNs to model the sequential relationships among the medical concepts, guided by future diagnoses prediction task in an end-to-end learning manner. Attention-based models, such as, MMORE [2] and MusaNet [5], have been employed to capture both visits' dependencies and sequential information in the EHR data to predict future diagnoses. Most of these methods rarely make use of the discharge states and irregular intervals in the EHR data. Recently, neural ODE [20], [21] has been proposed to handle arbitrary time gaps between observations, which provides an opportunity to alleviate the limitations of the existing models.

B. *Medical Ontologies on EHR*

Though healthcare information systems have accumulated huge volumes of EHR data, the data is generally not easily available due to both labour-intensive costs to label training data and privacy concerns of patient data. Facing the challenge of insufficient data, additional medical ontologies have been utilized to improve the quality of the medical code embeddings and the predictive performance. For instance, GRAM [3] proposes the graph-based attention model to incorporate the medical ontology with an attention mechanism and recurrent neural networks for representation learning with the application to diagnosis prediction. KAME [4] extends GRAM model to additionally consider side information of the learned embedding of the non-leaf nodes in medical ontology, and exploits RNN to integrate the knowledge from both the medical codes and non-leaf nodes and the EHR data to predict future diagnoses. MMORE [2] learns multiple ontological representations for the non-leaf nodes in the ontology and integrates the EHR co-occurrence statistics to predict sequential diagnoses. However, these models do not mutually integrate medical codes and the ontology, leaving learning effective representations from both EHR data and the ontologies an open question.

III. Methodology

This section starts with notations of several important concepts and problem statement in the paper. The remainder mainly focuses on details of the proposed model consisting of patient journey transformer, ontological and ODE representations, and task of sequential diagnoses prediction.

A. Notations and Problem Statement

1) Notations: We denote the set of medical codes from the EHR data as $c_1, c_2, \ldots, c_{|\mathbb{C}|} \in \mathbb{C}$ and $|\mathbb{C}|$ is the number of unique medical codes. Patients' clinical records can be represented by a sequence of visits $\boldsymbol{P} = \langle V_1, \ldots, V_t, \ldots, V_T \rangle$, which is referred to as the patient journey in the paper, where T is the number of visits in the patient journey. Each visit V_t consists of a subset of medical codes ($V_t \subseteq \mathbb{C}$). For clear demonstration, all algorithms will be presented with a single patient's journey. On the other hand, a medical ontology $\mathcal{G}$ contains the hierarchy of various medical concepts with the *parent-child* semantic relationship. In particular, the medical ontology $\mathcal{G}$ is a directed acyclic graph (DAG) and the nodes of $\mathcal{G}$ consist of leaves and their ancestors, shown in left part in Fig. 2. Each leaf node refers to a medical code in $\mathbb{C}$, which is associated with a sequence of ancestors from the leaf to the root of $\mathcal{G}$. And each ancestor node belongs to the set $\mathbb{N} = n_{|\mathbb{C}|+1}, n_{|\mathbb{C}|+2}, \ldots, n_{|\mathbb{C}|+|\mathbb{N}|}$, where $|\mathbb{N}|$ is the number of ancestor codes in $\mathcal{G}$. A ancestor node in the medical ontology $\mathcal{G}$ represents a related but more general concept over its children. Thus, including these semantic relationships would help the model to improve the medical concept representation that can lead to more accurate predictions of sequential diagnoses. Table I summarizes notations we will use throughout the paper.

TABLE I: Notations for SETOR.

Notation	Description
$\mathbb{C}$	Set of unique medical codes in dataset
$\lvert\mathbb{C}\rvert$	The number of unique medical codes
c_i	$c_i \in \mathbb{C}$, the i-th medical code in $\mathbb{C}, i = 1, \ldots, \lvert\mathbb{C}\rvert$
V_t	The t-th visit of the patient, $V_t \subseteq \mathbb{C}$
$\boldsymbol{P}$	The patient journey, $\boldsymbol{P} = \langle V_1, \ldots, V_t, \ldots, V_T \rangle$
$\mathcal{G}$	The medical ontology, a directed acyclic graph
$\mathbb{N}$	Set of ancestor codes in $\mathcal{G}$
$\boldsymbol{G}$	Ontological embedding matrix
$\boldsymbol{M}$	Embedding matrix of medical codes
$\boldsymbol{E}_{i,:}$	Basic embedding vector of medical code c_i
d	The dimension of medical code embedding

2) Problem Statement: Given a time-ordered patient journey $\boldsymbol{P} = \langle V_1, \ldots, V_t, \ldots, V_T \rangle$, and medical ontology $\mathcal{G}$, the goal of a sequential diagnosis prediction problem is to predict the next visit information. For the t-th visit, where $t = 1, 2, \ldots, T-1$, the outputs are $V_2, V_3, \ldots, V_T$.

B. Model Overview

To make the best use of the irregular and temporal properties in EHR and alleviate the challenge of insufficient data, we propose a robust and transformer-based model, called SETOR, illustrated in Fig. 2. First, the medical ontology $\mathcal{G}$ is embedded into an ontological embedding matrix $\boldsymbol{G}$. Then, an ontological encoder aggregates both the embedded diagnoses from $\boldsymbol{G}$ and an initial embedding matrix $\boldsymbol{M}$ by embedding operation $f(\cdot, \boldsymbol{P})$ to learn both co-occurrence and medical knowledge. $\boldsymbol{M} \in \mathbb{R}^{|\mathbb{C}| \times d}$ is an embedding matrix of medical codes, where d represents the embedding size. $\boldsymbol{M}$ is randomly initialised by a uniform distribution, and its entries are learnable during model training in an end-to-end manner. The outputs of the ontological encoder are fed into attention pooling layer to compress a set of medical codes in a visit into a vector representation. Next, our proposed LoS and Interval ODE representations are added to the learned visit representations, and the normalized outputs are fed into a journey transformer to learn the visit dependencies in a patient journey. Lastly, a predictive model, following the journey transformer, is used to predict the next visit information.

C. Ontological Representation

To mitigate the problem of data insufficiency in healthcare and to learn knowledgeable and generalized representations of medical codes, we employ the attention-based graph representation approach GRAM [3]. In the medical ontology $\mathcal{G}$, each leaf node c_i has a basic learnable embedding vector $\boldsymbol{E}_{i,:} \in \mathbb{R}^d$, where $1 \leq i \leq |\mathcal{C}|$, and d represents the dimensionality. And each non-leaf node n_i also has an embedding vector $\boldsymbol{E}_{i,:} \in \mathbb{R}^d$, where $|\mathbb{C}|+1 \leq i \leq |\mathbb{C}|+|\mathbb{N}|$. $\boldsymbol{E}$ is initialized with the values from the uniform distribution, and $\boldsymbol{E} \in \mathbb{R}^{(|\mathbb{C}| + |\mathbb{N}|) \times d}$. The attention-based graph embedding uses an attention mechanism to learn the d-dimensional final embedding $\boldsymbol{G}$ for each leaf node i (medical code) via:

$$\boldsymbol{G}_{i,:} = \sum_{j \in \, Pa_{\mathcal{G}}(i)} \alpha_{ij} \boldsymbol{E}_{j,:} \tag{1}$$

where $Pa_{\mathcal{G}}(i)$ denotes the set comprised of leaf node i and all its ancestors, $\boldsymbol{E}_{j,:}$ is the d-dimensional basic embedding of the node j and α_{ij} is the attention weight on the embedding $\boldsymbol{E}_{j,:}$ when calculating $\boldsymbol{G}_{i,:}$, which is formulated by following the Softmax function,

$$\alpha_{ij} = \frac{\exp(g(\boldsymbol{E}_{i,:}, \boldsymbol{E}_{j,:}))}{\sum_{k \in \, Pa_{\mathcal{G}}(i)} \exp(g(\boldsymbol{E}_{i,:}, \boldsymbol{E}_{k,:}))}. \tag{2}$$

$$g(\boldsymbol{E}_{i,:}, \boldsymbol{E}_{j,:}) = \boldsymbol{w}_\alpha^T \mathtt{tanh}\left(\boldsymbol{W}_\alpha(\boldsymbol{E}_{i,:} || \boldsymbol{E}_{j,:}) + \boldsymbol{b}_\alpha\right), \tag{3}$$

where $(\boldsymbol{E}_{i,:} || \boldsymbol{E}_{j,:})$ is to concatenate $\boldsymbol{E}_{i,:}$ and $\boldsymbol{E}_{j,:}$ in the child-ancestor order; $\boldsymbol{w}_\alpha$, $\boldsymbol{W}_\alpha$ and $\boldsymbol{b}_\alpha$ are learnable parameters.

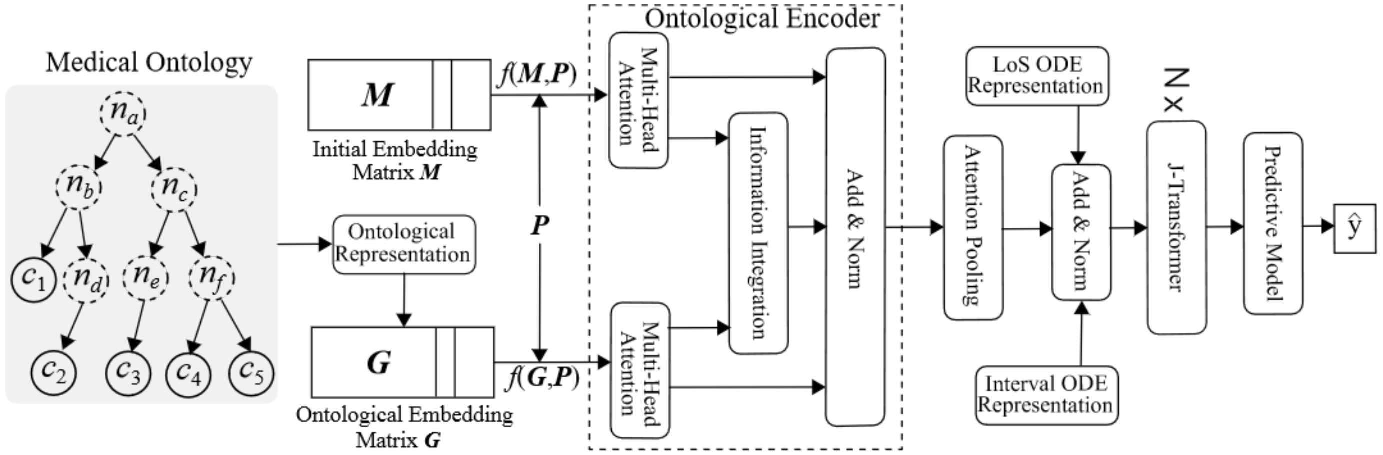


Fig. 2: The Proposed SETOR Model. The medical ontology is formatted as a directed acyclic graph, in which, the root node is virtual to construct the tree, the leaf nodes (solid circles) denote fine-grained diagnoses, and the non-leaf nodes (dotted circles) denote coarse-grained disease concepts.

D. Ontological Encoder

To encode both visit information and medical ontology as well as fuse their heterogeneous features, we propose ontological encoder. The dotted rectangle in Fig. 2 shows the details of the encoder, where. We first calculate the code embeddings $\mathbf{E}^M = f(\boldsymbol{M}, \boldsymbol{P})$ and node embedding $\mathbf{E}^G = f(\boldsymbol{G}, \boldsymbol{P})$, where $\mathbf{E}^M, \mathbf{E}^G \in \mathbb{R}^{(T-1)\times n \times d}$ are 3-dimensional tensors, the function f is to embed medical codes in patient journey $\boldsymbol{P}$ according to $\boldsymbol{M}$ and $\boldsymbol{G}$. Next, $\mathbf{E}^M, \mathbf{E}^G$ are fed into two different multi-head self-attentions (`MultiHead`) [27], where n is the number of medical codes in each visit of the patient journey. For simplicity, we take the t-th visit ($t = 1, 2, \ldots, T-1$) in patient journey $\boldsymbol{P}$ as an example to demonstrate the process of ontological encoder as follow,

$$\begin{aligned} \boldsymbol{V}_{Mt} &= \texttt{MultiHead}(\mathbf{E}^m_{t,:,:}, \mathbf{E}^m_{t,:,:}, \mathbf{E}^m_{t,:,:}), \\ \boldsymbol{V}_{Gt} &= \texttt{MultiHead}(\mathbf{E}^G_{t,:,:}, \mathbf{E}^G_{t,:,:}, \mathbf{E}^G_{t,:,:}). \end{aligned} \tag{4}$$

where `MultiHead` is a function of multi-head attention [27], and $\boldsymbol{V}_{Mt}, \boldsymbol{V}_{Gt} \in \mathbb{R}^{n\times d}$.

1) Multi-Head Attention: The multi-head attention mechanism relies on self-attention, where all of the keys, values and queries come from the same place. The self-attention operates on a query $\boldsymbol{Q}$, a key $\boldsymbol{K}$ and a value $\boldsymbol{V}$:

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}})\boldsymbol{V} \tag{5}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are $n \times d$ matrices, n denotes the number of medical codes in a visit in a patient journey, d denotes the dimension of embedding.

The multi-head attention mechanism obtains h (i.e. one per head) different representations of $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$, computes self-attention for each representation, concatenates the results. This can be expressed as follow:

$$\mathrm{head}_i = \mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V) \tag{6}$$

$$\mathrm{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Concat}(\mathrm{head}_1, ..., head_h)\boldsymbol{W}^O \tag{7}$$

where the projections are parameter matrices $\boldsymbol{W}_i^Q \in \mathbb{R}^{d\times d_k}, \boldsymbol{W}_i^K \in \mathbb{R}^{d\times d_k}, \boldsymbol{W}_i^V \in \mathbb{R}^{d\times d_v}$ and $\boldsymbol{W}^O \in \mathbb{R}^{hd_v\times d}$, $d_k = d_v = d/h$.

2) Information Integration: The ontological encoder adopts an information integration layer for the mutual integration of the code and node embedding in a visit. The process is as follows:

$$\boldsymbol{H}_t = \sigma(\boldsymbol{W}_M \boldsymbol{V}_{Mt} + \boldsymbol{W}_G \boldsymbol{V}_{Gt} + \boldsymbol{b}) \tag{8}$$

where $\boldsymbol{W}_M, \boldsymbol{W}_G, \boldsymbol{b}$ are learnable parameters, $\boldsymbol{H}_t \in \mathbb{R}^{n\times d}$ is the inner hidden state integrating the information of both the code and the node. $\sigma(\cdot)$ is the non-linear activation function, which usually is the ReLU function.

The output of ontological encoder is denoted as follows,

$$\boldsymbol{O}_t = \texttt{LayerNorm}(\boldsymbol{H}_t + \boldsymbol{V}_{Mt} + \boldsymbol{V}_{Gt}), \tag{9}$$

where $\boldsymbol{O}_t \in \mathbb{R}^{n\times d},\ 1 \le t \le (T-1)$, so that we can represent heterogeneous information of medical codes and ontology into a united feature space.

For ontological representation and encoder, we first build an embedding matrix for both leaf and non-leaf nodes in medical ontology. Then, we extract knowledge for each code from medical ontology as a tuple $(leaf,\ ancestors)$, and embed each code in the tuple by looking up the embedding matrix. Lastly, we calculate knowledge-enriched representation of each leaf node using Eq.(1). During this procedure, we also consider interactively encoding both visiting records and medical ontology by presenting an ontological encoder to fuse their heterogeneous features.

E. Attention Pooling

Attention Pooling [28], [29] explores the importance of each individual code within a visit. It works by compressing a set of medical code embeddings from the visit into a single context-aware vector representation. For simplicity, we take the t-th transformer output $\boldsymbol{V}^t$ as an example. Formally, it is written as:

$$g(\boldsymbol{V}^t_{i,:}) = \boldsymbol{w}^T \sigma(\boldsymbol{W}^{(1)} \boldsymbol{V}^t_{i,:} + b^{(1)}) + b, \tag{10}$$

where $\boldsymbol{V}^t_{i,:}$ is the i-th row of $\boldsymbol{V}^t$ ($1 \leq i \leq n$), σ is ReLU function and $\boldsymbol{w}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{b}$ are learnable parameters. The probability distribution is formalized as

$$\boldsymbol{\alpha}_t = \texttt{softmax}([g(\boldsymbol{V}^t_{i,:})]^n_{i=1}). \tag{11}$$

The final output $\boldsymbol{v}_t$ of the attention pooling is the weighted average of sampling a code according to its importance, i.e.,

$$\boldsymbol{v}_t = \sum_{i=1}^{n} \boldsymbol{\alpha}_t \odot [\boldsymbol{V}^t_{i,:}]^n_{i=1}, \tag{12}$$

where $\boldsymbol{v}_t \in \mathbb{R}^d$ $(1 \leq t \leq (T-1))$ represents the t-th visit in the patient journey.

F. ODE Representations

Neural ODE [20], [21], [30] models the time series as a continuously changing trajectory and makes better use of the data's timestamp information and predictions arbitrarily in time. Each trajectory is determined by the local initial state $\boldsymbol{h}_{t_1}$ and the global set of potential dynamics shared by the all-time series. Given observation time t_1, t_2,..., t_T and initial state $\boldsymbol{h}_{t_1}$, generated by the ODE solver $\boldsymbol{h}_{t_1}$,..., $\boldsymbol{h}_{t_{(T-1)}}$, which describes the underlying state of each observation. This generation model can be defined by the following formula:

$$\frac{\partial \boldsymbol{h}(t)}{\partial t} = f(\boldsymbol{h}(t), \boldsymbol{\theta}_f), \text{ where } \boldsymbol{h}(t_1) = \boldsymbol{h}_{t_1}, \tag{13}$$

$$\boldsymbol{h}_{t_1}, ..., \boldsymbol{h}_{t_{T-1}} = \texttt{ODESolve}\ (\boldsymbol{h}_{t_1}, f, \boldsymbol{\theta}_f, t_1, ..., t_{T-1}), \tag{14}$$

where f is a time-invariant function, using a neural network with parameters $\boldsymbol{\theta_f}$. The function takes the value $\boldsymbol{h}$ at the present time step and outputs the gradient at the end.

ODE is a function as $h(t)$ in Eq.(11). The equation needs to be solved during each evaluation and begins with an initial state h_{t_1}, which is also called initial value problem. Adjoint method by Pontryagin is employed to calculate the gradients of the ODE. With a low memory footprint, this method works by solving second, augmented ODE backwards in time and can be used with all ODE integrators. By solving the equation, the desirable sequence of hidden states can then be produced for downstream modules.

LoS ODE Representation. As each visit has two timestamps (admitted and discharge times) and each initial state (e.g. the i-th $\boldsymbol{v}_i, i = 1, .., (T-1)$) is given by the output of attention pooling, we can utilize Neural ODE to predict the discharge state as follow:

$$\boldsymbol{v}_i, \boldsymbol{v}_i^{dis} = \texttt{ODESolve}\ (\boldsymbol{v}_i, f, \boldsymbol{\theta}_f, t_i, t_i^{dis}), \tag{15}$$

where $\boldsymbol{v}_i^{dis}$ is the discharge state for the i-th visit, and t_i, t_i^{dis} are the admitted and discharge timestamps,respectively.

Interval ODE Representation. A patient journey consists of a sequence of irregular visits with timestamps. We can utilize Neural ODE to learn the hidden state for each visit timestamp following Equation 14.

$$\boldsymbol{h}_1 \sim p(\boldsymbol{v}_1), \tag{16}$$

$$\boldsymbol{h}_1, ..., \boldsymbol{h}_{T-1} = \texttt{ODESolve}\ (\boldsymbol{h}_1, f, \boldsymbol{\theta}_f, t_1, ..., t_{T-1}), \tag{17}$$

where p is a probability distribution dependent on time, and $\boldsymbol{v}_1$ is the first admitted visit state of a patient journey. The outputs of LoS and Interval ODE representations are added to the outputs of attention pooling and normalized for the following layer, which is denoted as follows:

$$\tilde{\boldsymbol{v}}_t = \texttt{LayerNorm}(\boldsymbol{v}_t + \boldsymbol{v}_t^{dis} + \boldsymbol{h}_t),\ 1 \leq t \leq (T-1). \tag{18}$$

G. Patient Journey Transformer Module

To learn visits' relationships in a patient journey, a module, called `J-Transformer`, is proposed to capture inherent dependencies, as shown in Fig. 2. `J-Transformer` responsible for learning dependencies of sequential visits in a patient journey with admission intervals and length of stay in each visit, which is calculated as follows:

$$\{\boldsymbol{v}_1^o, \ldots, \boldsymbol{v}_{T-1}^o\} = \texttt{J-Transformer}(\{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{T-1}\}). \tag{19}$$

Besides, we denote the number of `J-Transformer` layers as N. $\boldsymbol{o}_t$ is the input to `J-Transformer`, which is depicted in Fig. 3. `J-Transformer` is identical to its implementation in BERT [27] and [31], which has two sub-layers. The first is a multi-head attention mechanism mentioned in Section III-D1, and the second is position wise fully connected feed-forward network. Residual connection [32] is employed around each of the two sub-layers, followed by layer normalization [33].

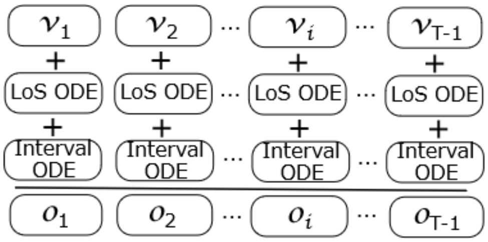


Fig. 3: Detailed inputs to `J-Transformer`

H. Sequential Diagnoses Prediction

Given a patient's visit records $\boldsymbol{P} = \{V_1, V_2, \ldots, V_{T-1}\}$, to capture the EHR sequential visit information we perform the sequential diagnoses predictive task with the objective of predicting the disease codes of the next visit V_t, which can be expressed as follows:

$$\hat{\boldsymbol{y}}_t = \texttt{Softmax}(\boldsymbol{W}\boldsymbol{v}_{t-1}^o + \boldsymbol{b}), \tag{20}$$

$$\mathcal{L} = \frac{1}{T}\sum_{t=2}^{T-1} -\left({\boldsymbol{y}_t}^{\top} \log \hat{\boldsymbol{y}}_t + (1-\boldsymbol{y}_t)^{\top} \log\left(1-\hat{\boldsymbol{y}}_t\right)\right), \tag{21}$$

where $\boldsymbol{v}_{t-1}^o \in \mathbb{R}^d$ is the output of `J-Transformer` to denote the representation of the $(t-1)$-th visit, $\mathcal{L}$ is the loss function, $\boldsymbol{y}_t$ is a vector with $|\mathbb{C}|$ elements, whose value is 1 if the i-th diagnosis code exists in V_t and 0 otherwise, and $\boldsymbol{W} \in \mathbb{R}^{|\mathbb{C}| \times d}$ and $\boldsymbol{b} \in \mathbb{R}^{|\mathbb{C}|}$ are the learnable parameters.

TABLE II: Statistics of the datasets.

Dataset	M-III	M-IV
# of patients	7,499	73,452
# of visits	19,911	295,351
Avg. # of visits per patient	2.66	4.02
# of unique ICD9 codes	4,880	9,165
Avg. # of ICD9 codes per visit	13.06	12.01
Max # of ICD9 codes per visit	39	57
# of category codes	272	283
Avg. # of cat. codes per visit	11.23	10.41
Max # of cat. codes per visit	34	37

IV. Experiments

In this section, we conduct experiments on two real world medical claim datasets to evaluate the performance of the proposed SETOR. Compared with the state-of-the-art predictive models, SETOR yields better performance on different evaluation strategies.

A. Data Description

We conducted comparative studies on two real-world datasets in the experiments, which are the MIMIC-III [34] and MIMIC-IV [35] databases.

a) M-III Dataset: The MIMIC-III dataset [34] is an open-source, de-identified dataset of ICU patients and their EHRs between 2001 and 2012. The diagnosis codes in the dataset follow the ICD9 standard. MIMIC-III is denoted by M-III in the experiment.

b) M-IV Dataset: The MIMIC-IV dataset [35] is an update to MIMIC-III, which incorporates contemporary data and improves on numerous aspects of MIMIC-III. The dataset consists of the medical records of 73,452 patients between 2008 and 2019. MIMIC-IV is denoted by M-IV.

Tab. II shows the statistical details about the datasets, where the selected patients made at least two visits. MIMIC-III and MIMIC-IV are denoted by M-III and M-IV in the experiment, respectively.

B. Experimental Setup

In this subsection, we first introduce the state-of-the-art approaches for diagnosis prediction task in healthcare, and then outline the measures used for predictive performance evaluation. Finally, we describe the implementation details.

a) Baseline Approaches: We compare the performance of our proposed model against the following state-of-the-art baseline models:

- **RETAIN** [11], which learns the medical concept embeddings and performs heart failure prediction via the reversed RNN with the attention mechanism.
- **Dipole** [10], which uses bidirectional RNN and three attention mechanisms (location-based, general, concatenation-based) to predict patient visit information. We chose location-based Dipole as a baseline method.
- **GRAM** [3], which is a graph-based attention model to learn the representations from the knowledge graph to predict future medical outcomes. .
- **KAME** [4], which is a diagnosis prediction model inspired by GRAM, using medical ontology to learn representations of medical codes and their parent codes. These are then used to learn input representations of patient data which are fed into a Neural Network architecture to predict sequential diagnoses.
- **MMORE** [2], which is based on medical ontology with an attention mechanism.

b) Predictive Task: The purpose of the sequential diagnosis prediction task is to predict the diagnosis information of the next visit. In the experiments, true labels $\boldsymbol{y}_t$ are prepared by grouping the ICD9 codes into 283 groups using CCS single-level diagnosis grouper[2]. It is to improve the training speed and predictive performance, while preserving sufficient granularity for all the diagnoses. We measure the predictive performance by $Accuracy@k$, which are defined as:

$$Accuracy@k = \frac{\text{\# of true positives in the top } k \text{ predictions}}{\text{\# of positives}}$$

Sequential diagnosis prediction is a multi-label problem, so normal Accuracy is inapplicable. Following previous works, we also use accuracy@k as the metric, which measures the ratio of positive labels ranked in top-k according to their logits. Specifically given a test sample, we first calculate the logits for all categories by a trained model, and rank the categories by the logits in descending order. Then, we count how many positive labels fall into top-k and compute the ratio over the number of all positive labels in this sample. Lastly, we derive the final metric, Accuracy@k, by averaging the ratios on the samples from the entire test set.

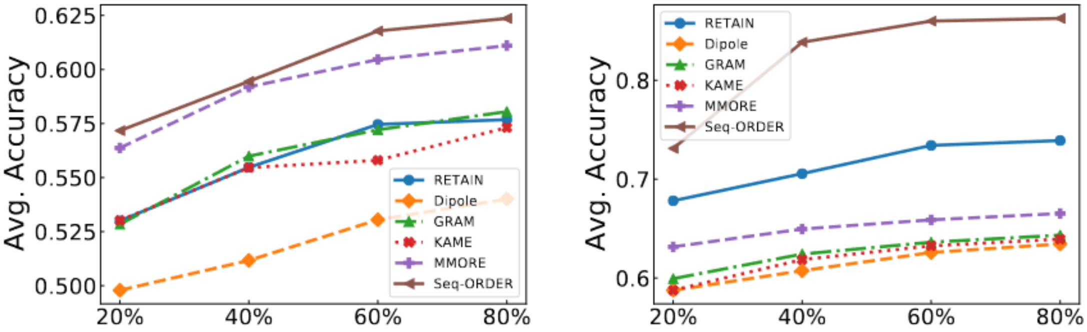
c) Implementation Details: We use CCS-multi-level diagnoses hierarchy as the medical ontology. We implement all the approaches with Pytorch 1.4.0. For the training models, we use Adadelta [36] with a minibatch of 32 patients. We randomly split the data into a training set, validation set and test set and fix the size of the validation set to be 10%. To validate the robustness against insufficient data, we vary the size of the training set from 20% to 80% and use the remaining part as the test set. The validation set is used to determine the best parameters values in the 100 training iterations. The drop-out strategies (the drop-out rate is 0.1) are used for all the approaches. We set dimension d = 200 for all the baselines and the proposed model.

C. Results of Sequential Diagnosis Prediction

Tab. III shows the accuracy@k of the SETOR and baselines with different k on two real-world datasets for the sequential diagnoses prediction task. From Tab. III, we can observe that the performance of the proposed SETOR is better than that of all the baselines on the two datasets.

On the M-III dataset, compared with the best baseline MMORE, the accuracy of SETOR improves by 2.21%. These results suggest that adding LoS and Interval ODE representation layers when predicting diagnoses is effective. We observe

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



(a) M-III

(b) M-IV

Fig. 4: Accuracy@20 of diagnoses prediction on M-III and M-IV, size of training data is varied from 20% to 80%.

that the performance obtained by the models using ontologies is better than that obtained by the models without using ontologies on M-III, which can be thought of as a small and insufficient dataset. The underlying reason is that the models integrating external medical ontologies can alleviate the issue of insufficient data.

On the M-IV dataset, the proposed SETOR still outperforms all the state-of-the-art diagnosis prediction approaches. Compared with the best baseline RETAIN, the accuracy of SETOR improves by 8.77% when $k = 5$. We also find that when not using ontologies RETAIN outperforms the other baseline models on M-IV the large dataset. This implies that the model can obtain comparative performance without using ontologies when the size of training dataset is larger. Compared to the models based on GRU and attention (e.g., RETAIN and Dipole), although Dipole fuses location attention, its performance is inferior to the attention-based models. Overall, our proposed framework exhibits better predictive power for both sufficient and insufficient datasets.

On the two datasets, the results show that the proposed model outperforms all the baselines, especially when the size of dataset is large. This demonstrates that the superiority of SETOR results from the explicit consideration of both the ontologies and the EHR co-occurrence, with the irregular intervals and discharge states being well handled.

D. Data Sufficiency Analysis

To analyze the influence of data sufficiency on the predictions, we randomly split the data into training, validation, and test sets, and fix the size of the validation set to be 10%. To validate the robustness against insufficient data, we vary the size of the training set to form four groups: 20%, 40%, 60%, and 80%, and use the remaining part as the test set. The training set in the 20% group are the most insufficient for training the proposed and baseline models, while the data in the 80% group are the most sufficient for training the models. Fig. 4 shows the Accuracy@20 on the M-III and M-IV.

From the Fig. 4a, we can observe that the accuracy of the proposed model is higher than that of the baselines in all groups. Specifically, MMORE is a comparative model with ontological representation and diagnosis co-occurrence over M-III, which shows that the models integrating medical ontology learns reasonable medical code embeddings to improve the prediction with insufficient data. The performance of Dipole is inferior to other baseline models, which indicates that the model only taking the sequential information into accounts is not enough.

TABLE III: Accuracy comparison of sequential diagnoses prediction.

Dataset	Model	Accuracy@k (%)			
		5	**10**	**20**	**30**
M-III	RETAIN	27.15	41.41	57.68	68.25
	Dipole	24.55	37.04	54.01	60.09
	GRAM	27.72	41.24	58.05	68.08
	KAME	27.98	41.81	57.31	68.02
	MMORE	28.97	43.74	61.10	71.61
	SETOR	**31.18**	**45.80**	**62.36**	**72.46**
M-IV	RETAIN	38.95	57.60	73.91	81.48
	Dipole	32.48	47.75	63.45	72.39
	GRAM	33.63	48.84	64.34	73.05
	KAME	33.56	48.80	63.94	72.64
	MMORE	34.21	50.59	66.53	75.21
	SETOR	**47.72**	**71.21**	**86.26**	**90.32**

When training data on the M-IV, which is a sufficient dataset, Fig. 4b shows that the proposed model significantly outperforms all the baselines. We observe that the performance obtained by RETAIN without using medical ontologies is better than that of the other baselines over M-IV. The underlying reason may be that the next-admission diagnosis prediction is more sensitive to the diagnosis co-occurrence and the sequential positions of the visits in sufficient data. Overall, the results demonstrate that the proposed model balances medical ontology and diagnosis co-occurrence over both insufficient and sufficient EHR data to further improve prediction performance.

E. Ablation Study

We performed a detailed ablation study to examine the contributions of the model's components to the prediction task. There are three components: (Transformer) the transformer blocks to learn the patient journey from the embedded vis-

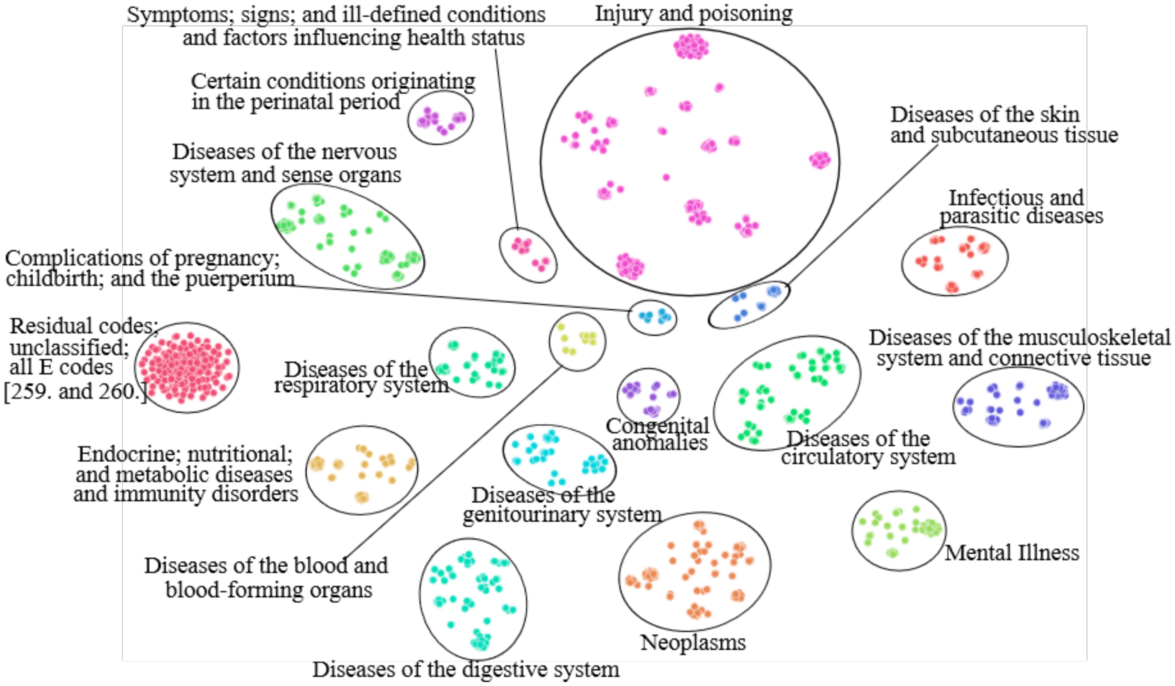


Fig. 5: Annotations of SETOR Diagnosis Embedding

its; (Ontology) the external medical ontology integrated into SETOR, and (ODE Representations) the LoS and interval encodings to be added to the learned visit embeddings.

- **w/o J-Trans:** remove the patient journey transformer blocks from the proposed model;
- **w/o Ontology:** remove the ontological representation from the proposed model;
- **w/o LoS:** remove LoS ODE representation;
- **w/o Interval:** remove the interval ODE representation;
- **w/o ODE:** replace two ODE representations with position embedding.

a) Ablated Transformers: `J-Transformer` is responsible for learning dependencies of sequential visits in a patient journey with admission intervals and length of stay in each visit. We conducted a group of experiments to analyze the contribution of this component to SETOR over two datasets. From Tab. IV, we observe that the full complement of SETOR achieved superior accuracy to the ablated models. Specifically, we note that the `J-Transformer` (w/o J-Trans) contributes the highest accuracy to the predictive task over M-III. Specifically, the accuracy improves by 4.14% and 4.89% when $k = 5$ and 20, respectively. On M-IV dataset, the performance of sequential diagnosis prediction is improved further with component of `J-Transformer`. The accuracy increases by 9.86% and 5.49% when $k = 5$ and 20, respectively. The underlying reason is that transformer blocks are providing significant amount of additional parameters and thus capacity. Thus, the prediction performance is significantly improved.

b) Ablated Representations: In the paper, representations consists of ontological, LoS and interval representation. From Tab. IV, we see that the components of various representations have contributions to the proposed model SETOR, though the contributions are no more than that of the `J-Transformer`. Specifically, we observe that the ontological representation (w/o Ontology) contributes the highest accuracy to the predictive task over M-III, which gives us confidence in using external medical ontologies to enhance the patient journey representations without sufficient data. Moreover, it is clear that the component of ODE representations provides valuable information for the performance of sequential diagnoses prediction over M-IV, which implies the irregular intervals and discharge states play more important roles with sufficient data. As shown in the ablation study (Tab. IV), the ontological and ODE representations contribute most to the predictive tasks, no matter the training data is sufficient or not, e.g., 1.18% lift of Acc@20 on M-IV and 0.56% lift of Acc@20 on M-III. Also as shown in Fig. 5, the embeddings produced by our proposed model shows the great inseparability of disease categories.

TABLE IV: Ablation Performance Comparison.

Ablation	**M-III** (%)		**M-IV** (%)	
	Acc@5	**Acc@20**	**Acc@5**	**Acc@20**
SETOR	**31.18**	**62.36**	**47.72**	**86.26**
w/o J-Trans	27.04	57.47	37.86	70.77
w/o Ontology	30.39	61.80	47.65	86.07
w/o LoS	30.74	61.96	47.59	86.20
w/o Interval	30.95	61.84	47.57	86.18
w/o ODE	30.45	61.93	47.38	85.08

F. Interpretable Representation Analysis

To qualitatively demonstrate the interpretability of the learned medical code embeddings by all the predictive models

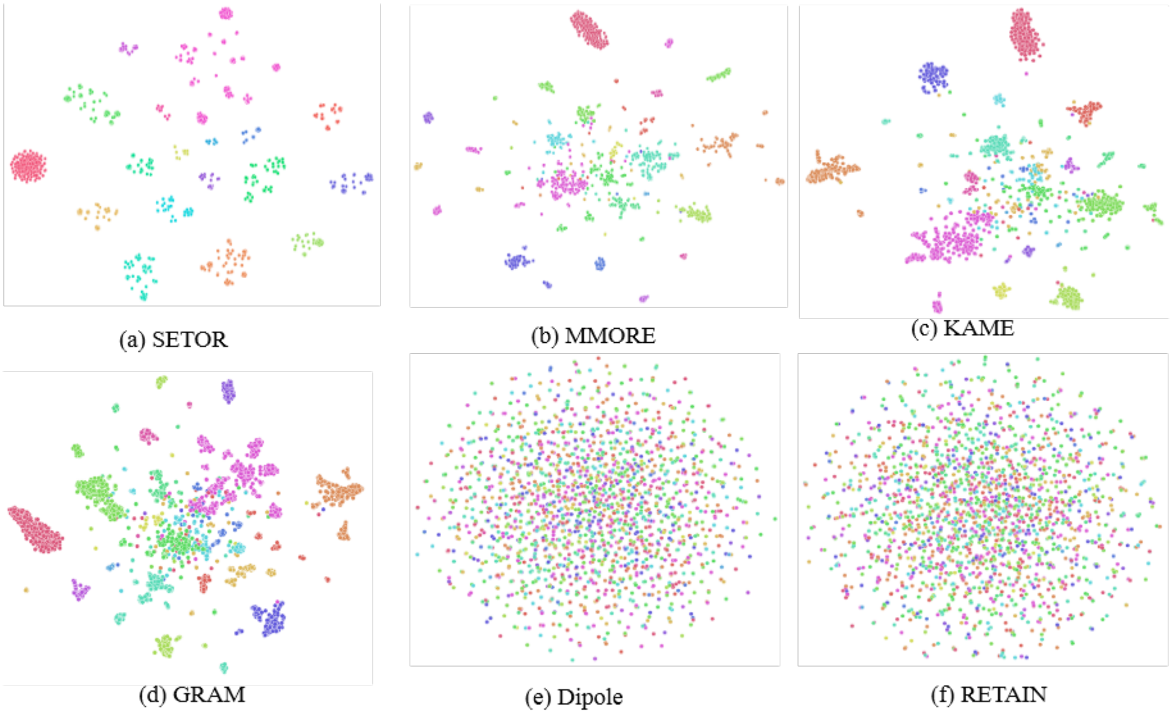


Fig. 6: *t*-SNE Scatterplots of Medical Codes Learned by Predictive Model on the M-III dataset.

on the M-III dataset, we randomly select 2000 medical codes and then plot on a 2-D space with t-SNE [37] as shown in Fig. 5 and Fig. 6. Each dot represents a diagnosis code, and the color of the dots represents the 18 disease categories in CCS multi-level hierarchy[3] and the text annotations represent the detailed disease categories. Ideally, the dots with the same color should be in the same cluster and there are margins among different clusters.

From Fig. 5, we can observe that SETOR learns interpretable disease representations that are in accord with the hierarchies of the given medical ontology $\mathcal{G}$, and obtains the 18 non-overlapping clusters. Specifically, for category of "Residual codes; unclassified; all E codes [259. and 260.]", we observe the medical codes in this category are closely clustered together, with large margin to other categories. The embedding results of medical codes in this category are harmony to CCS multi-level hierarchy, as category of "Residual codes; unclassified; all E codes [259. and 260.]" has not sub-category in CCS ontology. However, we note that the medical codes in category of "Injury and poisoning" are scattered in larger area. The underlying reason is that there are 12 sub-categories under this category. Thus, it is demonstrated that our proposed model SETOR learns meaningful and semantic representations for medical codes, which have practical interpretability.

As shown in Fig. 6, compared with SETOR, MMORE, KAME and GRAM are comparative baseline models, as those integrate medical ontology to predict sequential diagnoses. We observe the three baseline models learn reasonably interpretable diagnosis representations for several categories, as there is a large number of dots over-lapping in the center part of Fig. 6b, 6c, and 6c. It is clear that the medical codes in category of "Residual codes; unclassified; all E codes [259. and 260.]" are well clustered. But for the medical codes in category of "Injury and poisoning", their learned embeddings do not have clear margins to other categories. Fig. 6e and 6f suggest that models not using medical ontologies cannot easily learn interpretable representations. In addition, the predictive performance of SETOR is much better than that of MMORE, KAME, and GRAM shown in Tab. III, which proves that the proposed model does not affect the interpretability of medical codes. Moreover, it effectively improves the prediction accuracy.

V. Conclusion

Although the recent approaches have achieved promising performance on sequential diagnosis prediction task, they are still facing two major challenges, such as, how to effectively model irregular and temporal properties in EHR data and data insufficiency in healthcare information systems. In this paper, we propose an end-to-end transformer-based model, SETOR, to integrate medical ontology with visit information to mitigate the problem of data insufficiency, and to utilize neural ODE representations to learn hidden states for irregular intervals and visit discharges to effectively capture the irregular and temporal dependencies in EHR data. Although the proposed approach is focused on Electronic Health Record in healthcare domain, the core modules in this paper, e.g., ODE representations and ontological encoding, can be easily adapted into many other domains, such as, irregularly-sampled

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

time series. An experiment is conducted to show that SETOR outperforms baselines with both sufficient and insufficient data. The representations of medical codes are visualized to illustrate the interpretability of the proposed model. The experimental results on the two real-world medical datasets demonstrate the effectiveness, robustness, and interpretability of the proposed model.

Acknowledgment

This work was supported in part by the Australian Research Council (ARC) under Grant LP180100654 and DE190100626.

References

[1] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis," *IEEE J Biomed Health Inform*, vol. 22, no. 5, pp. 1589–1604, 2018.

[2] L. Song, C. W. Cheong, K. Yin, W. K. Cheung, B. C. Fung, and J. Poon, "Medical concept embedding with multiple ontological representations." in *IJCAI*, 2019, pp. 4613–4619.

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

[4] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "KAME: Knowledge-based attention model for diagnosis prediction in healthcare," in *CIKM*. ACM, Oct. 2018, pp. 743–752.

[5] X. Peng, G. Long, T. Shen, S. Wang, and J. Jiang, "Self-attention enhanced patient journey understanding in healthcare system," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2020, pp. 719–735.

[6] W. Chen, S. Wang, G. Long, L. Yao, Q. Z. Sheng, and X. Li, "Dynamic illness severity prediction via multi-task rnns for intensive care unit," in *ICDM*. IEEE, 2018, pp. 917–922.

[7] 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 *SIGKDD*. ACM, 2016, pp. 1495–1504.

[8] X. Zhang, B. Qian, X. Li, J. Wei, Y. Zheng, L. Song, and Q. Zheng, "An interpretable fast model for predicting the risk of heart failure," in *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 2019, pp. 576–584.

[9] X. Peng, G. Long, T. Shen, S. Wang, J. Jiang, and C. Zhang, "Bitenet: Bidirectional temporal encoder network to predict medical outcomes," in *ICDM*. IEEE, 2020, pp. 412–421.

[10] 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 *SIGKDD*. ACM, Aug. 2017, pp. 1903–1911.

[11] 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 *NeurIPS*, 2016, pp. 3504–3512.

[12] K. Jha, Y. Wang, G. Xun, and A. Zhang, "Interpretable word embeddings for medical domain," in *ICDM*. IEEE, 2018, pp. 1061–1066.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv:1301.3781*, 2013.

[14] Z. Qiao, S. Zhao, C. Xiao, X. Li, Y. Qin, and F. Wang, "Pairwise-ranking based collaborative recurrent neural networks for clinical event prediction," in *IJCAI*, 2018.

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

[16] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv:1406.1078*, 2014.

[17] G. Long, T. Shen, Y. Tan, L. Gerrard, A. Clarke, and J. Jiang, "Federated learning for privacy-preserving open innovation future on digital health," *arXiv preprint arXiv:2108.10761*, 2021.

[18] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "Ernie: Enhanced language representation with informative entities," *arXiv preprint arXiv:1905.07129*, 2019.

[19] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, "K-bert: Enabling language representation with knowledge graph." in *AAAI*, 2020, pp. 2901–2908.

[20] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *NeurIPS*, 2018, pp. 6571–6583.

[21] Y. Rubanova, R. T. Chen, and D. K. Duvenaud, "Latent ordinary differential equations for irregularly-sampled time series," in *NeurIPS*, 2019, pp. 5320–5330.

[22] E. Choi, C. Xiao, W. Stewart, and J. Sun, "Mime: Multilevel medical embedding of electronic health records for predictive healthcare," in *NeurIPS*, 2018, pp. 4547–4557.

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

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

[25] X. Peng, G. Long, S. Pan, J. Jiang, and Z. Niu, "Attentive dual embedding for understanding medical concepts in electronic health records," in *IJCNN*. IEEE, 2019, pp. 1–8.

[26] X. Peng, G. Long, T. Shen, S. Wang, J. Jiang, and M. Blumenstein, "Temporal self-attention network for medical concept embedding," in *ICDM*. IEEE, 2019, pp. 498–507.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[28] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv:1703.03130*, 2017.

[29] X. Cai, J. Gao, K. Y. Ngiam, B. C. Ooi, Y. Zhang, and X. Yuan, "Medical concept embedding with time-aware attention," in *IJCAI*, 2018, pp. 3984–3990.

[30] E. Dupont, A. Doucet, and Y. W. Teh, "Augmented neural odes," in *NeurIPS*, 2019, pp. 3140–3150.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[33] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[34] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[35] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "Mimic-iv (version 0.4)," *PhysioNet*, 2020.

[36] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[37] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.