Adaptive Integration of Categorical and Multi-relational Ontologies with EHR Data for Medical Concept Embedding

CHIN WANG CHEONG*, Hong Kong Baptist University, Hong Kong
KEJING YIN, Hong Kong Baptist University, Hong Kong
WILLIAM K. CHEUNG, Hong Kong Baptist University, Hong Kong
BENJAMIN C. M. FUNG, McGill University, Canada
JONATHAN POON, Hong Kong Hospital Authority, Hong Kong

Representation learning has been applied to Electronic Health Records (EHR) for medical concept embedding and the downstream predictive analytics tasks with promising results. Medical ontologies can also be integrated to guide the learning so that the embedding space can better align with existing medical knowledge. Yet, properly carrying out the integration is non-trivial. Medical concepts which are similar according to a medical ontology may not be necessarily close in the embedding space learned from the EHR data, as medical ontologies organize medical concepts for their own specific objectives. Any integration methodology without considering the underlying inconsistency will result in sub-optimal medical concept embedding, and in turn degrade the performance of the downstream tasks. In this paper, we propose a novel representation learning framework called ADORE (***AD**aptive **O**ntological **RE**presentations*) which allows the medical ontologies to adapt their structures for more robust integrating with the EHR data. ADORE first learns multiple embeddings for each category in the ontology via an attention mechanism. At the same time, it supports an adaptive integration of categorical and multi-relational ontologies in the embedding space using a category-aware graph attention network. We evaluate the performance of ADORE on a number of predictive analytics tasks using two EHR datasets. Our experimental results show that the medical concept embeddings obtained by ADORE can outperform the state-of-the-art methods for all the tasks. More importantly, it can result in clinically meaningful sub-categorization of the existing ontological categories and yield attention values which can further enhance the model interpretability.

CCS Concepts: • **Applied computing → Health informatics**.

Additional Key Words and Phrases: Electronic health record, representation learning, data mining with ontologies, predictive data analytics

ACM Reference Format:
Chin Wang Cheong, Kejing Yin, William K. Cheung, Benjamin C. M. Fung, and Jonathan Poon. 2018. Adaptive Integration of Categorical and Multi-relational Ontologies with EHR Data for Medical Concept Embedding. *ACM Trans. Intell. Syst. Technol.* 37, 4 (August 2018), 20 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding author: Chin Wang Cheong

Authors' addresses: Chin Wang Cheong, cwcheong@comp.hkbu.edu.hk, Hong Kong Baptist University, 224 Waterloo Road, Kowloon Tong, Hong Kong; Kejing Yin, cskjyin@comp.hkbu.edu.hk, Hong Kong Baptist University, 224 Waterloo Road, Kowloon Tong , Hong Kong; William K. Cheung, william@comp.hkbu.edu.hk, Hong Kong Baptist University, 224 Waterloo Road, Kowloon Tong , Hong Kong; Benjamin C. M. Fung, ben.fung@mcgill.ca, McGill University, 3661, Peel Street, QC H3A 1X1, Montreal, Quebec, Canada; Jonathan Poon, jonathan@ha.org.hk, Hong Kong Hospital Authority, 147B Argyle Street, Hong Kong.

2157-6904/2018/8-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Analyzing Electronic Health Records (EHR) data to achieve better patient care is attracting increasing attention. The representation learning approach has been found promising in analyzing the EHR data. It learns a latent embedding space to organize medical concepts where concepts frequently co-occurring in the EHR will be considered semantically similar and close in the embedding space [5, 9]. The learned embeddings could further be used as input for downstream applications, including clinical event prediction [6, 19, 28], next-admission diagnosis prediction [18, 24], treatment effect prediction[16], among others. Despite the promising results obtained, the inevitable noise and missing information makes it challenging to learn highly accurate representation. In addition, the semantic relationship among the medical concepts captured by the embedding space may not well align with the existing medical knowledge. This further limits the reliability and interpretability of the approach.

Integrating medical ontologies for medical concept embedding has been explored to enhance the alignment between the embedding space and the medical knowledge. For example, Clinical Classifications Software (CCS) which encodes the hierarchical categorization of diseases has been adopted for integrating into the representation learning process [6, 19, 31]. The integration generally leads to better organization of the embedding space in that the relationship between medical concepts align with the given ontology. Yet, there also exists *inconsistency* between the co-occurrence information in the EHR data and the semantic relationship described in the ontologies. For instance, according to the CCS ontology, the two diagnoses "Type I Diabetes Mellitus (T1DM, ICD-9 code: 25001)" and "Type II Diabetes Mellitus (T2DM, ICD–9 code: 25000)" share the same ancestor "Diabetes mellitus without complication" (**is-a** relationship) are supposed to be close in the medical concept space. However, it is inconsistent to the fact that they never co-occur in the EHR data as they are mutually exclusive. It is desired that the learned embeddings could respect such information in the EHR data in addition to following the structure of the ontologies. Finding the optimal way for the integration of medical ontologies and EHR data to achieve better representation learning and thus improve prediction accuracy is still an open issue.

In addition, there are multiple publicly available medical ontologies which are designed for different purposes and hence organize the medical concepts differently. How to utilize these resources together and at the same time minimize their potential *inconsistency* is another important and yet challenging problem. The CCS is a *categorical ontology* with one relationship type (is-a). The Systematized Nomenclature of Medicine Clinical Terms (SNOMED–CT) is a *multi-relational ontology* where rich medical knowledge is specified using multiple types of relationships. For instance, the diagnosis "Fracture of leg" has the relationship **associated-morphology** with "Fracture", and "Fracture of leg" also has the relationship **finding-site** with "Bone structure of lower limb". Such relationship triplets (entity–relationship–entity) obviously provide additional information for medical concept representation learning. While we can learn for each relationship type an embedding for integrating the categorical ontology CCS and the multi-relational ontology SNOMED-CT in the embedding space in ways similar to the knowledge graph embedding approach, the optimality of such integration is yet to be studied.

To address the issues, we argue that it is vital to allow more adaptive integration of multiple ontologies with the EHR data so that the ontologies can have their ways of organizing the medical concepts adjusted "collaboratively" under the representation learning framework. Based on our preliminary work on representation learning from categorical ontology and the EHR data [31], we further extend the idea to learn an adaptive number of basic embeddings for each category so that medical concepts under the same category can readily be reorganized to be not *necessarily* close in the embedding space unless the EHR data supports that. An attention mechanism similar to the idea of "multi-sense" for word embedding [13] is introduced to allow adaptive node splitting (except for the root node) to achieve the automatic ontology reorganization. Take diabetes mellitus again as the example, "T1DM" and "T2DM" are very different in the sense of therapy while they rarely co-occur in the data set. Therefore, they may be represented by two different "senses" of their shared ancestor. Furthermore, for integrating as well multi-relational ontologies, we first learn each relationship type an embedding and propose a *category-aware*

graph attention network where the category information according to the categorical ontology is taken into account to obtain the embedding of the relationship.

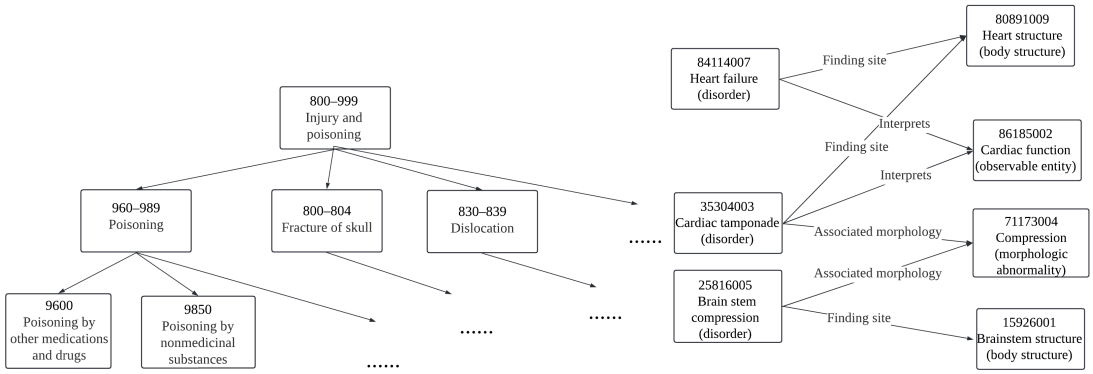
In this paper, we propose ***AD**aptive **O**ntological **RE**presentations (ADORE)*, that incorporates the aforementioned mechanisms for adaptive integration of the EHR data, the categorical ontologies and the multi-relational ontologies. We evaluated the performance of ADORE by integrating the categorical ontology CCS and the multi-relational ontology SNOMED–CT with two different open-source EHR datasets — MIMIC–III [15] and eICU [26]. Our experimental results demonstrate that ADORE could learn representations that not only align with the medical ontologies, but also respect the EHR data; therefore, it has much improved interpretability compared to existing methods. In addition, the boost in predictive performance also validates the effectiveness of ADORE and the obtained representations. To the best of our knowledge, incorporation of multi-relational ontologies into medical concept representation learning has not yet been well explored. Also, this is the first attempt that allows adaptive integration of the categorical ontology CCS, the multi-relational ontology SNOMED–CT and the EHR data to achieve robust and interpretable medical concept representation learning where the inconsistency issue is explicitly addressed.

2 RELATED WORK

Representation learning has been actively explored for medical predictive analysis. The earlier work mostly adopts one-hot representation, and the semantic relationships cannot be well preserved [30]. The vector-based representation can also be learned in ways similar to Word2Vec [21] by considering the co-occurrence information of the medical concepts in the EHR data [5] or clinical narratives [9]. Similarly, Daehr [35] which is an extension of Linear Discriminant Analysis (LDA) exploits the disease co-occurrence to detect mental health disorders. Models like RETAIN [7], Dipole [18], and MiME [8] also consider the ordering of the co-occurrence of the clinical events and learn recurrent neural networks to capture the sequential relationships among the medical concepts. In addition, there are works that apply representation learning to the EHR data in the federated learning setting [10].

In addition to the EHR data, publicly available medical ontologies also provide rich information for learning the medical concept embedding space. The International Statistical Classification of Diseases and Related Health Problems (ICD) [33] is a widely used disease coding system where diseases and procedures are classified hierarchically. The Anatomical Therapeutic Chemical (ATC) [34] Classification System is a drug classification system of active ingredients of drugs according to their therapeutic, pharmacological, and chemical properties. The Systematized Nomenclature of Medicine Clinical Terms (SNOMED–CT) [23] is a standard used in U.S. Federal Government systems for the electronic exchange of clinical health information, which consists of medical concepts including diseases, drugs and their relationships. Medical ontologies are essentially knowledge graphs. Various knowledge graph embedding methods have been developed and found effective when applied to real-world knowledge graphs like Wordnet [22] and Freebase [3]. In particular, the translation-based family (e.g., TransE [4], TransH [32], TransD [14], TransR [17], among others [12, 37]) has been rigorously studied where entities and relations of knowledge graphs are projected onto the low-dimensional embedding spaces for applications like completion of the missing relations. Alternatively, the Relational Graph Convolutional Network (R-GCN) [27] has also been proposed for modeling multi-relational data.

Approaches for integrating EHR data with categorical medical ontologies (e.g., CCS) using deep learning models for medical concept embedding have been explored in the literature. GRAM [6] adopts a graph-based attention model to learn the embedding space from ontologies. KAME [19] makes use of a knowledge attention mechanism based on the patient representation to attend to the medical concepts in the knowledge graph. HAP [38] tries to propagate attention across the whole ontology hierarchically so that a medical concept learns its embedding with reference to all other medical concepts in the hierarchy instead of only its ancestors. G-BERT [29] enhances the medical concept embedding space learned from the medical ontology using BERT and graph neural networks.



(a) CCS (b) SNOMED–CT

Fig. 1. Snapshots of the medical ontologies used in this paper. (a) The CCS is a categorical ontology, providing hierarchical information for diagnoses. (b) The SNOMED–CT is a multi-relational ontology that describes different types of relationships between medical concepts and entities.

Some other variations include the adoption of Transformer (e.g., SETOR [25]) and leveraging information extracted from the clinical notes (e.g., Trove [11]).

Yet, integrating the EHR data and the ontologies for medical concept embedding is still an open research question. First, the aforementioned models do not explicitly consider the potential inconsistency between the co-occurrence patterns learned from the EHR data and the semantic similarity described in the ontologies. Also, extending them to include also multi-relational ontologies (like SNOMED–CT) is non-trivial. ADORE is proposed in this paper to address the issues by allowing more *adaptive* integration.

3 NOTATIONS AND PRELIMINARIES

3.1 Basic Notations

We denote the medical concepts in the EHR data set such as diagnoses and medication codes by $c_1, c_2, \ldots, c_{|C|} \in C$, where $|C|$ is the total number of medical concepts in the given EHR data. The hospital admissions of a patient are denoted as $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_t$, where each contains a subset of the medical concepts, *i.e.* $\mathcal{A}_t \subset C$ for the t^{th} hospital admission. Also, the hospital admission $\mathcal{A}_t$ is denoted by a binary vector $\mathbf{x}_t \in \{0,1\}^{|C|}$ where the i^{th} entry of $\mathbf{x}_t$ equals to one if $c_i \in \mathcal{A}_t$, and zero otherwise.

In the paper, we focus on the problem of integrating the EHR data with a categorical medical ontology and a multi-relational medical ontology for predictive analytics. We denote the medical concept embeddings corresponding to the categorical ontology as $\mathbf{V}$ and the ones corresponding to the multi-relational ontology as $\mathbf{G}$. Notations used in this paper are listed in Table 1.

3.2 Medical Ontologies

Medical ontologies can be represented as knowledge graphs with the entities corresponding to *medical concepts* (e.g., diagnosis codes, body parts, etc.) and the relations corresponding to the relationship between them. They are created with different purposes and thus take different structures. *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD–9–CM) [33] is a categorical medical ontology that was developed for hierarchical categorization of diseases. It follows a tree structure, where the leaf nodes represent specific diseases and their ancestors are disease categories. E.g., 921.3 “contusion of eyeball” is under the category of

921 "contusion of eye and adnexa". *The Clinical Classifications Software* (CCS)[1] for ICD-9-CM is a diagnosis and procedure categorization scheme where the ICD-9-CM's multitude of codes are collapsed into a smaller number of clinically meaningful categories that are useful for presenting descriptive statistics. Fig. 1(a) illustrates a small portion of the CCS ontology. SNOMED–CT [23] is a multi-relational medical ontology, which is a standardized, multilingual vocabulary of clinical terminology. It contains relationships of different types to link up medical concepts, e.g., diagnoses, drugs, and body parts. Figure 1(b) shows a small snapshot of SNOMED–CT where three different disorders are connected to some medical concepts via different relationships.

Table 1. Notations used in this paper.

Notation	Description
c_i	The i^{th} medical concept
C	The set of all medical concepts
$\mathcal{A}_t$	The t^{th} hospital admission
$x_t \in \{0,1\}^{\|C\|}$	The binary representation of $\mathcal{A}_t$
y_t	The label of the t^{th} hospital admission
$\mathbf{e}_j^i$	The i^{th} basic embedding of entity j in categorical ontology
$\mathbf{h}_k$	The basic embedding of entity k in multi-relational ontology
$\mathbf{V}$	Medical concept embedding matrix (categorical ontology)
$\mathbf{G}$	Medical concept embedding matrix (multi-relational ontology)
$\mathbf{W}$	Medical concept embedding matrix (co-occurrence in EHR)
$\mathbf{U}$	Final medical concept embedding matrix

4 PROPOSED MODEL

In this paper, we propose a novel representation learning framework ADORE to learn the medical concept embedding space. ADORE allows a categorical medical ontology and a multi-relational medical ontology to *adaptively* reorganize their structures and representations for better integration with the EHR data in the embedding space. ADORE contains four components: 1) a multiple ontological representation network to learn the embeddings of medical concepts according to a tree-based categorical ontology (e.g., CCS), 2) a category-aware graph attention network (CAGAT) that can derive a flexible representation of the relationships between different medical concepts from a multi-relational ontology (e.g., SNOMED–CT) for more adaptive integration, 3) a co-occurrence embedding module that utilizes the co-occurrence statistics from the given EHR database, and 4) a prediction network that drives the learning of the overall embedding space. Fig. 2 depicts the overview of the proposed ADORE framework.

4.1 Learning Multiple Embeddings in Categorical Ontology

A categorical ontology (e.g., CCS) can be represented in the form of a tree with its leaf nodes corresponding to the set of all medical concepts C. As in [6, 31], the "ancestor-descendant" relationships among the nodes of the tree are used to learn the embeddings of the medical concepts. Specifically, a node in the tree is associated with a *basic embedding* $\mathbf{e} \in \mathbb{R}^d$. The *final embedding* of a node is calculated from the convex combination of attention coefficients from node's ancestors.

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

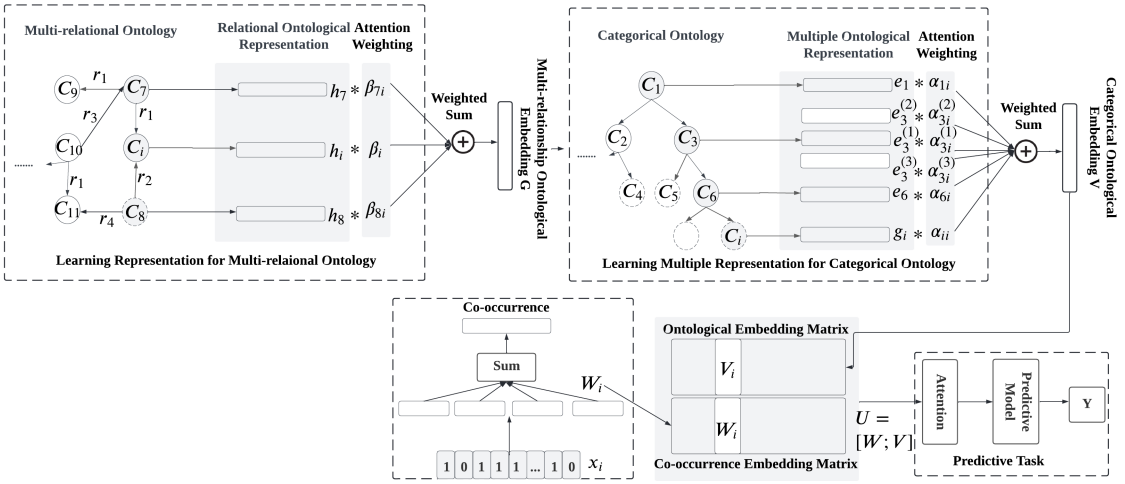
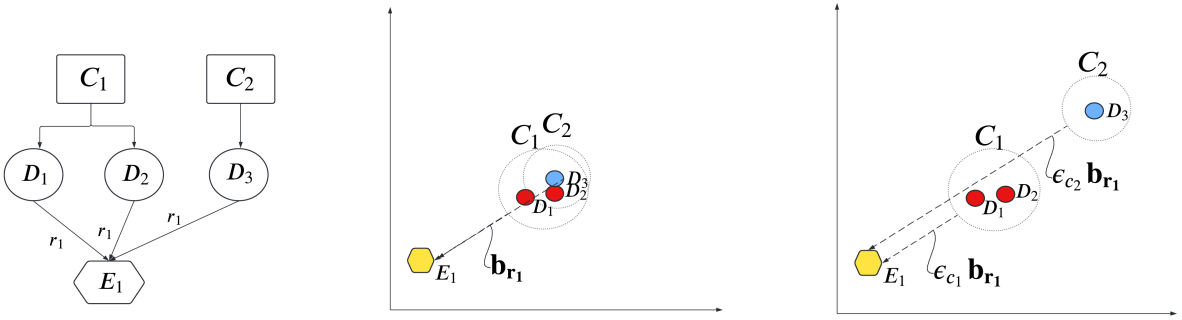


Fig. 2. **Overview of the proposed ADORE framework.** ADORE integrates EHR data with multi-relational ontologies and categorical ontologies for learning a medical concept embedding space. The upper left shows the category-aware multi-relational graph attention network that derives the multi-relational embeddings **G**. The upper right shows the multiple ontological representation network that learns multiple ontological representations for the categorical ontology with **G** as its input. The leaf nodes (dotted circles) indicate the medical concepts, while the non-leaf nodes (solid circles) denote the categories in ontologies. The basic embeddings of the nodes in the categorical ontology (upper right) and the multi-relational ontology (upper left) are denoted as **e** and **h** respectively. For the categorical ontology, ADORE first learns multiple basic embeddings for each node (except the root node), and then combines them via an attention mechanism to derive the categorical medical concept embeddings **V**. The final overall embeddings **U** are formed by the concatenation of **V** and co-occurrence embeddings **W** for computing the predictive result.



(a) (b) (c)

Fig. 3. An illustration of the adaptive integration of categorical and multi-relational ontologies to learn the embedding space. (a) A simple categorical ontology with three diseases D_1, D_2 and D_3 belonging to two different categories C_1 and C_2, and connected to the same medical entity E_1 via relationship r_1 according to a multi-relational ontology. (b) The medical concept embeddings learned by the GCN methods where diseases of categories C_1 and C_2 are undesirably placed together due to their shared relationship with E_1. (c) ADORE alleviates this limitation by introducing the category-aware relational bias $\epsilon_{c_i}\mathbf{b}_r$ to separate the disease embeddings corresponding to the two categories.

The validity of this methodology assumes that the embedding of each node to be similar to its ancestor's, which may not be supported by the observations in the EHR data. Simply enforcing that will result in inconsistency between the knowledge in the ontology and patterns discovered in the EHR data. We therefore propose to allow the categories, i.e., non-leaf nodes (except the root node) in the tree, to carry multiple semantic meanings, or "senses". Formally, instead of learning one single vector representation for each non-leaf node, multiple basic embeddings, denoted as $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots$, are assigned. For each category node c_j, we estimate the optimal number of the basic embeddings $K(j)$ adaptively. Similar to [31], we adopt an attention mechanism for learning the multiple basic embeddings for the nodes in the tree so that the final embedding $\mathbf{v}_i$ for node c_i can be derived as:

$$\mathbf{v}_i = \alpha_{ii}\mathbf{g}_i + \sum_{j \in \text{path}(i)} \sum_{k=1,\ldots,K(j)} \left(\alpha_{ji}^{(k)} \mathbf{e}_j^{(k)}\right), \tag{1}$$

where $\mathbf{g}_i$ is the basic embedding of leaf node c_i, $\mathbf{e}_j^{(k)}$ is the k^{th} basic embedding for the category node c_j, $K(j)$ is the number of basic embeddings of node c_j, and α is the attention weightings which are non-negative and sum up to one, *i.e.* $\sum_{j \in \text{path(i)}} \sum_{k=1,\ldots,K(j)} \alpha_{ji}^{(k)} = 1, \alpha_{ji}^{(k)} \geq 0$. Note that we fix $K(1) = 1$ for the root node so that the root node has only one basic embedding. To obtain the attention weightings α, we compute the score between the basic embedding of a category $\mathbf{e}_j$ and that of a leaf node $\mathbf{g}_i$ via a compatibility function, given by:

$$f(\mathbf{e}_j, \mathbf{g}_i) = \mathbf{s}^\intercal \tanh\left(\mathbf{M} \begin{bmatrix} \mathbf{e}_j \\ \mathbf{g}_i \end{bmatrix}\right), \tag{2}$$

where $\mathbf{s}$ and $\mathbf{M}$ are both learnable parameters. Then, we compute the attention weightings by using the softmax function, *i.e.*,

$$\alpha_{ji}^{(k)} = \frac{\exp(f(\mathbf{e}_j^{(k)}, \mathbf{g}_i))}{\sum_{j' \in \text{path}(i)} \sum_{k'=1,\ldots,K(j)} \left(\exp(f(\mathbf{e}_{j'}^{(k')}, \mathbf{g}_i))\right)}. \tag{3}$$

The idea of using multiple embeddings per category has been shown promising in [31], where the value of $K(j)$ was simply fixed to be two. Here we argue that for a medical concept (i.e., a leaf node) c_i, if its ancestor (i.e., a non-leaf node) indexed by j has attention score $\alpha_{ji}^{(k)}$ that is sufficiently low, it will have very marginal impact on the final embedding of the medical concept c_i. Therefore, it is less likely to force the final embedding of c_j to be close to its sibling. Therefore, we need only increase the number of basic embeddings $K(j)$ for the category node c_j with a higher attention score (say larger than a threshold). With this insight, we initialize the number of basic embeddings of all non-leaf nodes to be one, and adaptively increase $K(j)$ for all the category nodes during the training process by:

$$K_{new}(j) \leftarrow \begin{cases} K_{old}(j) + 1 & \text{if } \max_i(\alpha_{jiK_{old}(j)}) > \lambda, \\ K_{old}(j) & \text{otherwise,} \end{cases} \tag{4}$$

where λ is a pre-defined threshold on the attention score. We tested different values for λ and found that it works best when it is set to 0.8.

Take the CCS class *Other bacterial pneumonia* as an example, ADORE can adaptively learn three basic embeddings to represent three subgroups of this class. Specifically, three pneumonia diseases related to streptococcus (48230, 48241, and 48242) are represented by the first embedding, the pneumonia due to pseudomonas (4821) and hemophilus (4822) by the second embedding, and the unspecified pneumonia by the third embedding. Those adaptive splits of embedding representations are clinically meaningful and so provide better sub-grouping of diseases. On the other hand, the CCS class *Empyema and pneumothorax* contains only two diseases (Primary spontaneous pneumothorax and secondary spontaneous pneumothorax). Forcing it to have multiple embeddings, as in [31], is unnecessary because these two diseases are clinically similar. The primary one indicates pneumothorax in the absence of known lung disease, and the secondary one is for the patients with underlying lung diseases.

Instead, ADORE sticks with only one embedding for this class during the training to promote the parameter sharing and learn a more compact space.

4.2 Learning Category-aware Relational Embeddings in Multi-relational Ontology

There are publicly available multi-relational medical ontologies which provide information about the relationships among the medical concepts [23]. Incorporating them into the representation learning framework [4] in principle can guide the learning to achieve higher accuracy. Yet, different ontologies are established for specific purposes with distinct semantic relationships encoded. Integrating them by simply combining them, say fusing a categorical ontology and a multi-relational ontology into a single graph, could inevitably end up with conflicting objectives for guiding the representation learning. Fig. 3 shows an example to illustrate this point. In the example, there are two diseases D_1 and D_2 under the category C_1, and another one D_3 under the category C_2 according to a categorical ontology. At the same time, all three diseases are related to a clinical entity E_1 with r_1 being the relationship according to another multi-relational ontology. As D_1, D_2, and D_3 all share the relationship with a common entity, their embeddings would naturally be close in the embedding space, as shown in Fig. 3(b). This however is undesirable as they are under different categories according to the categorical ontology and their embeddings should be sufficiently distinct.

To allow it to respect also the information in the categorical ontology, we propose a category-aware graph attention network (CAGAT) to learn embeddings for a multi-relational ontology (e.g., SNOMED–CT) with reference to the categorical heterogeneity encoded in the categorical ontology (e.g., CCS). The key idea is that CAGAT allows a length-varying relational bias to be learned so that the embeddings of diseases under two different categories can stay apart in the embedding space even they share the same type of relationship with the same entity, as illustrated in Fig. 3(c). Specifically, for each medical concept, we first compute a new embedding based on the multi-relational ontology and then use it to replace the basic embedding of the corresponding leaf-node for the categorical embedding learning in Eq. ((1)). We denote the embedding of the node corresponding to the i^{th} medical concept in the multi-relational ontology as $\{\mathbf{h}_i\}$. For each node, a weighted sum of its own embedding and those of its neighbors in the multi-relational ontology can be computed with an attention mechanism, given as:

$$\mathbf{g}_i = \beta_{ii}\mathbf{h}_i + \sum_{(j,r)\in N_i} \beta_{ij}(\mathbf{h}_j + \epsilon_i \mathbf{b}_r), \tag{5}$$

where N_i is the set of all node-relationship pairs associated with $\mathbf{h}_i$, $\mathbf{b}_r$ is the learnable relationship bias for the r^{th} relationship, and ϵ_i is the category-wise coefficient determining how far $\mathbf{h}_i$ and $\mathbf{h}_j$ should fall apart, which in turn depends on the category of the i^{th} medical concept in the categorical ontology. The ϵ_i is estimated as:

$$\epsilon_i = \sum_{j\in \text{path}(i)} \frac{\mathbf{p}^\top \mathbf{s}_j}{\left\| \sum_{j'\in \text{path}(i)} \mathbf{s}_{j'} \right\|_2}, \tag{6}$$

where path(i) is the set of ancestors of the medical concept c_i in the categorical ontology, $\mathbf{s}_j$ is the auxiliary embedding of the category indexed by j and $\mathbf{p} \in \mathcal{R}^D$ is a learnable parameter that projects the embedding onto a scalar.

Then, the attention weightings can be obtained by applying the softmax function, *i.e.*,

$$\beta_{ij} = \frac{\exp(g(\mathbf{e}_j + \epsilon_i \mathbf{b}_\mathbf{r}, \mathbf{g}_i))}{\sum_{(k,t)\in N_i} \exp(g(\mathbf{e}_k + \epsilon_i \mathbf{b}_\mathbf{t}, \mathbf{g}_i))}, \tag{7}$$

where the function $g(\cdot,\cdot)$ is approximated using a single layer perceptron, *i.e.*,

$$g(\mathbf{e}_j, \mathbf{e}_i) = \mathbf{n}^\intercal \tanh\left(\mathbf{H}\begin{bmatrix} \mathbf{e}_j \\ \mathbf{e}_i \end{bmatrix} + \mathbf{z}\right), \tag{8}$$

and $\mathbf{n}$, $\mathbf{z}$ and $\mathbf{H}$ are the parameters to be learned.

Considering the implementation, the proposed CAGAT contains much fewer parameters to be learned in the multi-relational attention layer, compared with the ordinary relational graph convolutional network (R-GCN) [27]. The latter needs $O(R \times D^2)$ parameters for the weighting matrices where R is the number of relationships and D is the dimension of the embeddings. CAGAT consists of only $O(R \times D)$ parameters, and thus is less prone to over-fitting.

4.3 Learning Embeddings by Co-occurrence Statistics

We can also learn the medical concept embeddings based on the EHR data. Given a hospital admission $\mathcal{A}_t$ that contains a set of l medical concepts $\{c_1, c_2 ... c_l\} \in C$, we first calculate the average of their embeddings as the "context" of the admission, *i.e.*, $\mathbf{a}_t = (\sum_{k=1}^{l} \mathbf{w}_k)/l$. The key idea of learning embedding from the co-occurrence statistics is that the "context" should be able to predict the medical concepts which may exist in the hospital admission $\mathcal{A}_t$. That can be implemented by minimizing the negative log-probability of the concepts existed in the admission conditioned on the "context", *i.e.*,

$$\mathcal{L}_t^{\text{co-occur}} = -\frac{1}{l}\sum_{k=1}^{l} \log p(c_k|\mathbf{a}_t) = -\frac{1}{l}\sum_{k=1}^{l} \log \frac{\exp(\mathbf{w}_k'^T \mathbf{a}_t)}{\sum_{i=1}^{l} \exp(\mathbf{w}_i'^T \mathbf{a}_t)}, \tag{9}$$

where $\mathbf{w}'$ are learnable parameters. And the conditional probability is provided by the output of the softmax function.

4.4 Interpretability-Enhanced Predictive Analytics

Finally, we added the predictive task in the framework, which offers supervised information to the ontological representations. Specifically, we use the embeddings learned from the ontologies and the EHR data to predict the next-admission diagnosis, mortality, or re-admission risk. At the same time, the attention mechanism is exploited to further enhance the interpretability and the prediction. The representation matrices $\mathbf{W}$ and $\mathbf{V}$ are first concatenated row-wise as the final representations of medical concepts, *i.e.*, $\mathbf{U} = [\mathbf{W}; \mathbf{V}]$. For a hospital admission $\mathcal{A}_t$ denoted as a binary vector $\mathbf{x}_t$, we first compute the intermediate embedding for the hospital admission by retrieving the representations from $\mathbf{U}$ and summing them up, *i.e.* $\mathbf{a}_t = \mathbf{U}\mathbf{x}_t$. Then we calculate the attention-based admission representation $\tilde{\mathbf{a}}_t$ as follows:

$$\tilde{\mathbf{a}} = \sum_{i=1}^{|\mathcal{A}|} \beta_i * \mathbf{u}_i, \quad \sum_{i=1}^{|\mathcal{A}|} \beta_i = 1, \ \beta_i \geq 0, \tag{10}$$

where we ignore the subscript t denoting the t^{th} hospital admission for simplification of the notations, β is the attention weighting for the predictive task, and i indicate the index of the medical concept presented in the admission $\mathcal{A}$. The attention weighting is calculated by the softmax function:

$$\beta_i = \frac{\exp(g(\mathbf{a}, \mathbf{u}_i))}{\sum_{k=1}^{|\mathcal{A}|} \exp(g(\mathbf{a}, \mathbf{u}_k))}, \tag{11}$$

where we approximate the function $g(\cdot, \cdot)$ by a single layer perceptron with the same form as Eq. (2).

After we derive the attention-based admission embedding $\tilde{\mathbf{a}}_t$, it is used as input to the prediction model:

$$\hat{\mathbf{y}}_t = \text{softmax}\left(\tanh\left(\mathbf{Q}\tilde{\mathbf{a}}_t + \mathbf{k}\right)\right), \tag{12}$$

where $\mathbf{Q} \in \mathbb{R}^{d \times |N|}$ and $\mathbf{k} \in \mathbb{R}^{|N|}$ are the learnable parameters, d is the dimension of the final admission embedding and $|N|$ is the number of classes of output labels. We formulate the cross-entropy loss as the objective function

as follows:

$$\mathcal{L}_p^{\text{pred}} = -\frac{1}{T-1}\sum_{t=1}^{T-1}\left[\mathbf{y}_t^\intercal \log(\hat{\mathbf{y}}_t) + (1-\mathbf{y}_t)^\intercal \log(1-\hat{\mathbf{y}}_t)\right], \tag{13}$$

where T is the number of hospital admissions of the p^{th} patient, and $\mathbf{y}_t$ is the ground truth label.

By combining the objective functions of the predictive task and that of the EHR co-occurrence (Eq. 9), and taking average over all the patients, we can derive the overall objective function as follows:

$$\mathcal{L} = \frac{1}{N_p}\sum_{p=1}^{N_p}\left(\mathcal{L}_p^{\text{pred}} + \frac{1}{T_p}\sum_{t=1}^{T_p}\mathcal{L}_t^{\text{co-occur}}\right) \tag{14}$$

where T_p is the number of admissions of the p^{th} patient, and N_p is the total number of patients.

5 EXPERIMENTS

To evaluate the proposed ADORE[2], we carry out quantitative and qualitative evaluation using real-world large-scale datasets MIMIC-III and eICU. For the prediction performance, we conduct the next-admission diagnosis prediction, mortality prediction, and re-admission prediction. In addition, we demonstrate the interpretability of the embeddings learned through case studies. Besides, we present some particular attention weightings derived by ADORE with discussions on their quality.

5.1 Data and Ontology Pre-processing

We use MIMIC-III (Medical Information Mart for Intensive Care) [15] and eICU [26] datasets for evaluation. The former is a public dataset containing data of over 46,000 patients admitted to intensive care units (ICU) of the Beth Israel Deaconess Medical Center between 2001 and 2012. The latter is a multi-center database, containing over 200,000 ICU admissions across the United States between 2014-2015.

For the categorical medical ontology, we adopt CCS[3]. For the multi-relational medical ontology, we adopt SNOMED–CT[4] which consists of diagnoses, medications and additional entities such as body-site (*e.g.*, "head", "face", and so on). Relationships include **due-to**, **following**, **finding-site**, etc. The relationships link entities together to form a multi-relational graph.

5.1.1 Data Pre-processing. We filter out the patients with less than two hospital visits for the next diagnosis task and we exclude hospital admissions that have no diagnosis and medication records. We also remove the base type medications for MIMIC-III, *e.g.*, D5W.

For MIMIC-III, we finally extract 6, 453 patients with 2.7 hospital admissions per patient on average. The average numbers of diagnoses and medications in each admission are 12.0 and 38.9, respectively. For the eICU, we extract 8, 565 patients which contains 27, 635 hospital admissions. The average numbers of diagnoses and medications in each admission are 14.65 and 21.47, respectively.

5.1.2 SNOMED–CT Pre-processing. Relationships such as *Has presentation strength denominator value* and *Count of base of active ingredient* are removed as ADORE only makes use of discrete labels. In addition, *is-a* relationship is removed as it is repeated with the CCS ontology. After the pre-processing, there remain 23 different types of relationships. They include *Has definitional manifestation*, *Direct morphology*, *Access instrument*, *Has focus*, *Following*, *Specimen substance*, *Has precise active ingredient*, *Interprets*, *Subject relationship context*, *Approach*, *Using substance*, *Associated procedure*, *Has active ingredient*, *Has measured component*, *Associated finding*, *Finding method*,

[2]The code is available on Github at https://github.com/KenCheong/Adaptive-Integration-of-Categorical-and-Multi-relational-Ontologies-with-EHR-Data-for-Medical-Concep.git

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

[4]http://www.snomed.org/

Table 2. Statistics of categories with different numbers of embedding splits at each CCS level for CCS.

CCS Level	# of categories	Average # of basic embeddings
1	20	2.75
2	136	1.97
3	367	1.59
4	209	1.44

Temporally follows, *Has BoSS, Access*, *Method*, *Direct substance*, *Using access device, Due to*. Note that SNOMED–CT provides an official mapping[5] between ICD–9 codes and the concepts in SNOMED–CT, we use this mapping to ensure that the annotations of the overlapping concepts are consistent.

5.2 Baseline Models

We compare the performance of ADORE against the following models proposed for medical concept embedding:

- *Med2Vec* [5] which exploits sequential and co-occurrence relationships of medical concepts.
- *GRAM* [6] which exploits the categorical ontology with an attention mechanism and was applied to diagnosis prediction.
- *HAP* [38] which propagates attention across the entire ontology so that the medical concept embeddings are learned from all concepts in the hierarchy instead of only ancestors. As it is designed for the next-admission diagnosis prediction, we only show the results of the method for the task.
- *MMORE* [31] which combines the categorical ontology and the EHR data by allowing two basic embeddings for each non-leaf node to improve the flexibility.
- *ADORE-cat*, which is a variant of ADORE that removes the multi-relational ontology component.
- *ADORE*, which is a full version of ADORE that includes both the categorical ontology and the multi-relational ontology components.

5.3 Experiment Setup

We determine the hyperparameters by the grid search. We set the dimension of the ontological embedding (both categorical and multi-relational) and the co-occurrence embedding to be 400 in our model. The embedding dimension is set as 800 for all the baselines. The dimension of the hidden layer for the attention mechanism is set as 100. The model is optimized using Adadelta [36] with the batch size of 100. And the maximum number of basic embeddings for the entities in the categorical ontology is set as 3.

5.4 Predictive Evaluation Tasks

We evaluate the predictive performance using several important clinical prediction tasks, including the next-admission diagnosis prediction [18, 24], mortality prediction [28], and hospital re-admission prediction[2]. We randomly split the data into the training set, test set, and validation set. We fix the size of the validation set to be 10% of the dataset. To evaluate the robustness against different settings of insufficient data, we vary the size of the training set from 20% to 80% of the dataset and use the remaining part as the test set.

For the next-admission diagnosis prediction, we follow a similar approach as in [6, 31] and derive the ground-truth labels $\mathbf{y}_t$ for diagnoses prediction by grouping the diagnoses in the next admissions into 712 groups based on the first three digits of their ICD-9 codes and conduct multi-label classification. We measure the performance

[5]https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html

Table 3. Performance comparison on MIMIC-III in terms of Accuracy@20 for next-admission diagnosis prediction, and AUC for mortality prediction and re-admission prediction. The result is averaged over 5 runs and the comparison is carried out over training sets of different sizes. The entries with significant improvement based on significant test ($P \leq 0.05$) are marked with *.

Model	10%	20%	40%	80%
Re-admission prediction				
Med2Vec	0.5564±0.0035	0.5796±0.0022	0.5868±0.0023	0.6248±0.0014
GRAM	0.6515±0.0039	0.6491±0.0036	0.6753±0.0032	0.6996±0.0013
MMORE	0.6588±0.0042	0.6632±0.0033	0.6818±0.0025	0.6972±0.0019
ADORE-cat	0.6523±0.0042	**0.6693±0.0034**	0.6840±0.0032	0.6988±0.0025
ADORE	**0.6635±0.0045**	0.6667±0.0035	**0.6914±0.0021***	**0.7176±0.0023***
Next-admission diagnosis prediction				
Med2Vec	0.4912±0.0027	0.4963±0.0031	0.5187±0.0019	0.5290±0.0015
GRAM	0.4974±0.0029	0.5058±0.0025	0.5315 ±0.0025	0.5550±0.0011
HAP	0.5021±0.0032	0.5214±0.0031	0.5318 ±0.0023	0.5510±0.0018
MMORE	0.4958±0.0030	0.5176±0.0029	0.5383±0.0028	0.5621±0.0012
ADORE-cat	0.4985±0.0035	0.5168±0.0026	0.5398±0.0022	0.5668±0.0009
ADORE	**0.5023±0.0033**	**0.5229±0.0025***	**0.5466±0.0011***	**0.5695±0.0019***
Mortality prediction				
Med2Vec	0.8213±0.0017	0.8632±0.0013	0.9078±0.0011	0.9341±0.0008
GRAM	0.8833±0.0016	0.9010±0.0017	0.9132±0.0014	0.9421±0.0015
MMORE	0.8820±0.0019	0.9148±0.0017	0.9241±0.0018	0.9460±0.0011
ADORE-cat	0.8822±0.0020	0.9237±0.0022	0.9269±0.0017	0.9525±0.0014
ADORE	**0.8882±0.0021**	**0.9322±0.0016***	**0.9273±0.0013***	**0.9529±0.0015***

of next-admission diagnosis prediction using *Accuracy@k*, which is defined as:

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

For mortality prediction and hospital re-admission prediction, we use the area under the ROC curve (AUC) as the evaluation metric as both are binary classification tasks. Due to highly imbalanced labels for the mortality prediction task, we retain all patients with positive labels (patients deceased in hospital) and randomly sample the same number of patients with negative labels for the mortality prediction task. For the hospital re-admission prediction task, we predict whether each patient will be admitted to hospitals in the future given the current visit.

5.5 Predictive Performance

The results of the predictive tasks are summarized in Tables 3 and 4 for MIMIC–III and eICU datasets, respectively. The results show that ADORE consistently achieve the best predictive performance in all the tasks conducted on both datasets, demonstrating the effectiveness of ADORE. Compared with Med2Vec, all other methods achieve significant improvement in predictive performance, especially when the training size is small. This indicates that integrating medical ontologies for learning the medical concept embedding could help alleviate the issue of insufficient data. Although the same categorical ontology is being used, MMORE outperforms GRAM for most of the training sizes and datasets. Take the next-admission diagnosis prediction task as an example. Using only 20% data for training, MMORE can improve the Accuracy@20 from 0.5058 to 0.5176 and from 0.6229 to 0.6423 for MIMIC–III and eICU respectively. This is because GRAM does not attempt to resolve the inconsistency between

Table 4. Performance comparison on eICU in terms of Accuracy@20 for next-admission diagnosis prediction, AUC for mortality prediction, and Re-admission prediction based on training sets of different sizes.

Model	10%	20%	40%	80%
Re-admission prediction				
Med2Vec	0.5232±0.0026	0.5286±0.0025	0.5324±0.0022	0.5466±0.0019
GRAM	0.5361±0.0030	0.5431±0.0026	0.5562±0.0023	0.5527±0.0022
MMORE	0.5376±0.0033	0.5445±0.0031	0.5585±0.0030	0.5563±0.0025
ADORE-cat	0.5300±0.0035	0.5453±0.0029	0.5580±0.0032	0.5607±0.0024
ADORE	**0.5501±0.0033***	**0.5471±0.0027**	**0.5634±0.0028***	**0.5662±0.0025***
Next-admission diagnosis prediction				
Med2Vec	0.5064±0.0027	0.5747±0.0028	0.6540±0.0022	0.7053±0.0011
GRAM	0.6065±0.0029	0.6229±0.0019	0.6749±0.0018	0.7161±0.0010
HAP	0.5523±0.0024	0.6456±0.0025	0.6792±0.0037	0.7208±0.0012
MMORE	0.6057±0.0023	0.6423±0.0021	0.6716±0.0019	0.7108±0.0011
ADORE-cat	0.5985±0.0024	0.6439±0.0026	0.6784±0.0022	0.7192±0.0014
ADORE	**0.6112±0.0029***	**0.6457±0.0026**	**0.6796±0.0015**	**0.7226±0.0010***
Mortality prediction				
Med2Vec	0.5812±0.0029	0.6344 ±0.0021	0.6512±0.0016	0.6569±0.0012
GRAM	0.5916±0.0031	0.6529±0.0029	0.6614±0.0021	0.6644±0.0013
MMORE	0.6237±0.0032	0.6455±0.0030	0.6777±0.0022	0.6882±0.0015
ADORE-cat	**0.6397±0.0035**	0.6429±0.0027	0.6826±0.0023	0.6931±0.0019
ADORE	0.6363±0.0035	**0.6815±0.0029***	**0.7016±0.0025***	**0.7160±0.0018***

the EHR data and the medical ontology. MMORE, on the other hand, learns two basic representations for each category in the ontology to provide the flexibility for better alignment. The performance of HAP in the next diagnosis prediction is comparable to MMORE as it tries to improve the limitation of GRAM by propagating information to more ancestors for diseases. ADORE-cat which estimates the number of basic embeddings needed for each category can further improve the predictive performance. ADORE with all components included gives the best prediction performance. It achieves significant improvement in terms of predictive performance over GRAM, especially when the training size is small. For instance, when the training size is as small as 10%, ADORE outperforms GRAM by 7.6% and 2.6% for mortality and re-admission prediction tasks, respectively, based on the eICU dataset. This implies that ADORE is effective in achieving adaptive integration of the two different types of medical ontologies and the EHR data within a unified representation learning framework.

5.6 Evaluating the quality of the ontological embeddings

To quantitatively evaluate the organization of similar concepts in the learned ontological embedding space, we employ the evaluation method proposed in [1]. In particular, we calculate the intrinsic measures of the quality of the embedding space by calculating the average distance in the latent space between nodes that are close in the ontologies. As the ontologies we are considering include both categorical and multi-relational, we propose two different scores, namely *categorical score* and *relational score* as follows.

Table 5. Performance comparison on ontological embedding quality based on intrinsic measures.

Model	Categorical	Categorical-Single	Relational	Joint
Med2Vec	0.2624	0.2563	0.4872	0.4168
GRAM	**0.9517**	**0.9378**	0.5512	0.7322
MMORE	0.9476	0.9145	0.6033	0.7622
ADORE-cat	0.8522	0.9114	0.5912	0.8129
ADORE	0.6689	0.9273	**0.9274**	**0.8403**

$$\textit{Categorical score} = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{1}{|\text{path}(i)|}\sum_{j\in\text{path}(i)}\mathcal{S}\left(h_i, \frac{1}{|\text{ch}(j)|}\sum_{k\in\text{ch}(j)}h_k\right) \tag{15}$$

$$\textit{Relational score} = \frac{1}{|C|}\sum_{i=1}^{|C|}\sum_{(j,r)\in N_i}\frac{1}{|N_r'|\times R}\mathcal{S}\left(h_i, \frac{1}{|N_j|}\sum_{(k,r')\in N_j}h_k\right) \tag{16}$$

where $\mathcal{S}(\cdot,\cdot)$ denotes the cosine similarity between two embedding vectors in the latent space, $\text{ch}(\cdot)$ denotes the set of children nodes of the j^{th} node, N_i denotes the set of node-relationship pairs associated with the i^{th} node, $|N_r'|$ is the number of nodes that connects to at least one other node with the r^{th} relationship, and R is the number of types of relationships.

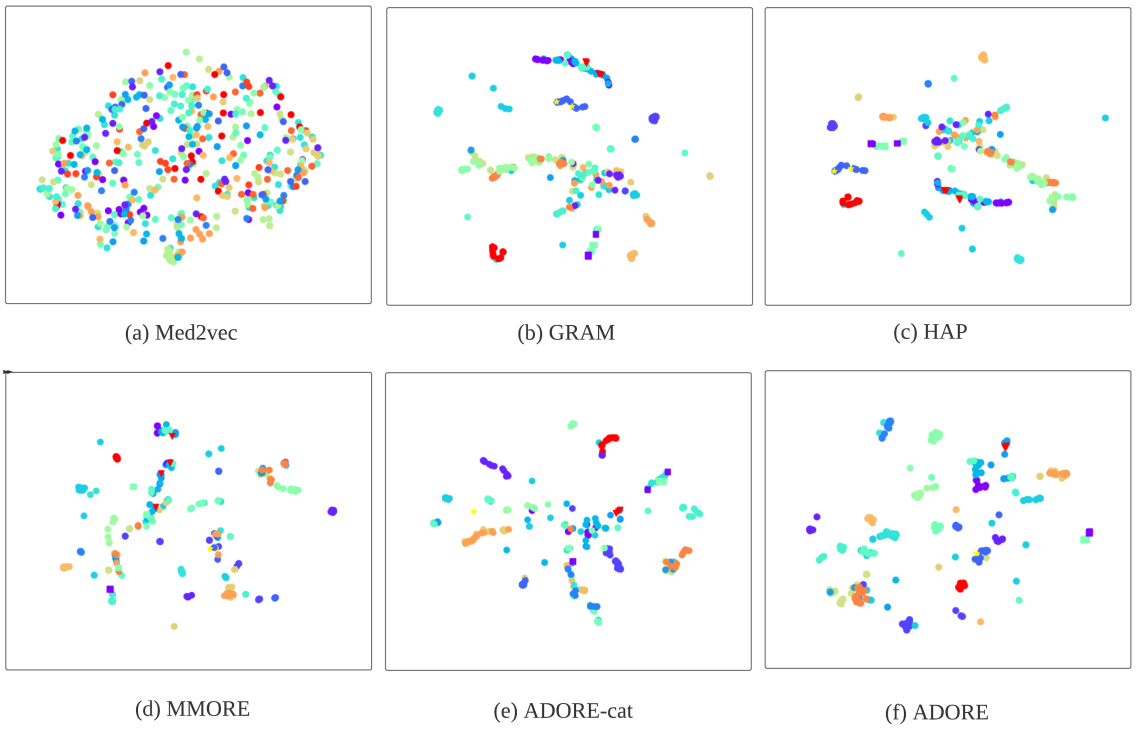
Also, to measure the overall performance with both categorical and multi-relational information considered at the same time, we also compute a joint score using Eq. (16) with the parent-children relationship in the CCS ontology being considered as one of the relationships.

The results of comparison are summarized in Table 5. When we consider only the categorical score, ADORE has a lower value than the other methods like GRAM as expected. This is because ADORE allows each non-leaf node to carry multiple semantic meanings. Intuitively, the increased number of splits for each non-leaf node will encourage more diverse medical concept embeddings if supported by the EHR data, even if they belong to the same category in the ontology. To verify this, we further compare only the non-leaf nodes that were not split during the training (i.e., ADORE infers that such nodes require only one basic embedding for its representation) and report the categorical score for those nodes in the column "Categorical-Single". ADORE obtains a similar performance to that obtained by the others. The result implies that for the medical concepts which are similar as revealed in the EHR data and described in the CCS ontology, the embeddings learned by ADORE in fact can align well with the ontology.

With respect to the relational and joint scores, ADORE significantly outperforms the other models, indicating that the embeddings learned by ADORE not only addresses the inconsistency between the EHR data and the CCS ontology, but also properly respect the relational SNOMED-CT ontology.

5.7 Interpretability of Embeddings Learned

To compare the interpretability of ADORE with other methods, we show an overview of final representations of diagnosis codes learned by different models using t-SNE [20] in Fig. 4. The different colors indicate the CCS categories of the diagnosis code. We randomly select 50 categories from the third level counting from the bottom in the ontology (excluding the leaf level). Fig. 4(a) is obtained by Med2Vec, which learns the embedding using the EHR data alone. By incorporating the CCS ontology, GRAM and HAP, as visualized in Fig. 4(b) and (c), could learn the embeddings with some clustering structure. Yet, the middle part of the visualization shows that medical codes under different categories cannot be well disentangled. The inconsistency between the EHR and



(a) Med2vec (b) GRAM (c) HAP

(d) MMORE (e) ADORE-cat (f) ADORE

Fig. 4. Scatter diagrams of the embeddings learned by different methods. Nodes from 50 randomly selected lowest level categories in the CCS ontology are visualized. And the colors refer to different CCS categories.

medical ontologies not being addressed is believed to be the reason. MMORE and ADORE-cat, as shown in Fig. 4(d) and (e), respectively, could give embeddings with clustering structures much more prominent than GRAM and HAP. This implies that allowing multiple basic embeddings to be learned could allow a more adaptive integration to help better alleviate the inconsistency issue, and thus lead to medical embeddings that could better capture the semantic meanings of the medical codes. Fig. 4(f) shows the medical concept embeddings learned by ADORE. It is evident that ADORE gives more distinct clustering structure for the representations learned than all other methods, implying that properly integrating the multi-relational ontology as proposed could further help capture the latent semantic relationships between different medical codes, and thus improve both the predictive performance and the interpretability.

5.7.1 Case Study. To further demonstrate the effectiveness of ADORE by incorporating both categorical and multi-relational ontologies, we conduct a case study to examine the embeddings learned for several diagnoses. The results are summarized in Fig. 6 and the relationship between the diagnoses examined are shown in Fig. 5.

We extract diagnosis codes under four CCS categories: *Viral pneumonia*, *Other bacterial pneumonia*, *Pulmonary tuberculosis* and *Regional enteritis*. In the SNOMET–CT ontology, the categories *Viral pneumonia* and *Other bacterial pneumonia* link to the entity *Lung structure* with the relationship of *finding site*. The category *Regional enteritis* links to *Granulomatous inflammation* with the relationship of *associated morphology*. The category *Pulmonary tuberculosis* links to both SNOMED–CT entities at the same time.

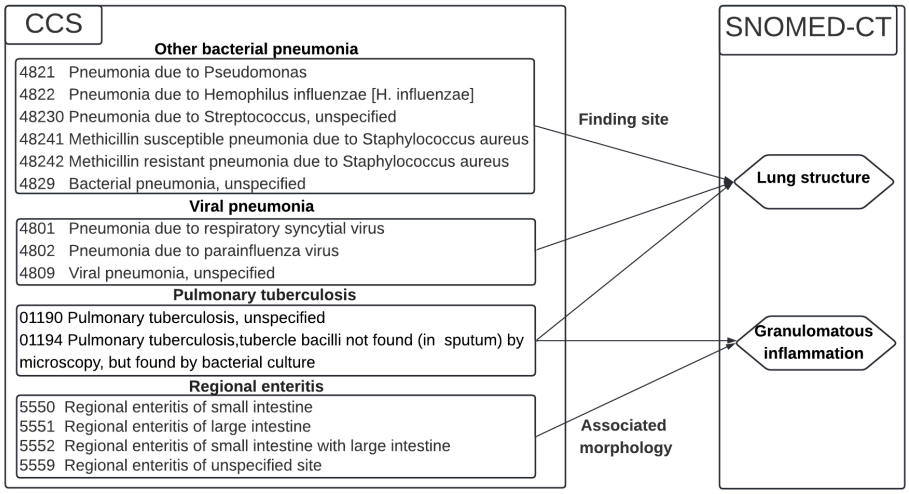


Fig. 5. A sub-graph with entities and relationships extracted from CCS and SNOMED–CT for case study.

As shown in Fig. 6(a), GRAM groups diagnosis codes under *Viral pneumonia* and *Other bacterial pneumonia* together because they share the same ancestor *Pneumonia and influenza* in the CCS ontology. MMORE, as shown in Figure 6(b), introduces additional embedding representation for the classes and thus two classes are separated in the latent space. However, as the number of embeddings allowed for each category in MMORE is fixed to two, the setting may not be optimal for all categories. For example, the diagnosis codes under *Regional enteritis* are separated into two groups in the space where 5550 belongs to one group, and the remaining ones belong to the other one. This separation does not align with the existing medical knowledge, and thus is not desirable. On the other hand, ADORE-cat in the Figure 6(c) can discover the subgroups of *Pneumonia and influenza* while grouping the diagnosis codes under *Regional enteritis* together. The reason is that we allow the number of additional embeddings to be adaptive to each category during training. Table 2 shows the statistics of the number of splits learned for the categories at different levels of the ontology based on MIMIC-III. We can observe that categories in higher levels tend to split more than the lower ones.

Figure 6(d) shows the embeddings learned by ADORE, which further incorporates the multi-relational ontology SNOMED–CT. Compared to other models, ADORE learns a well-organized embedding space that forms a clear hierarchy of medical concepts. For diagnosis codes under *Regional enteritis*, 5559 is organized to be far away from the other three diagnosis codes as it is an unspecified site code. Besides, *Viral pneumonia* and *Other bacterial pneumonia* are clearly separated into two distinct groups in the space but still closer than the other diagnosis codes due to the use of the category-aware relational bias. For the *Viral pneumonia*, 4809 is distant from 4801 and 4802 as 4809 is an unspecified subtype of viral pneumonia. When zooming in the *Other bacterial pneumonia*, we can observe two subgroups. One group consists of three five-digit codes (48230, 48241, and 48242) which are related to streptococcus. For the other group consisting of three four-digit codes (4821, 4822, 4829), it can be further divided into two: 4829, which is an unspecified code, and 4821, 4822. Meanwhile, the embeddings of diagnosis codes under *Pulmonary tuberculosis*, via the learning, is organized in the middle region of diseases that link to either *Lung structure* or *Granulomatous inflammation* because they share both connections.

5.7.2 Interpretation of the Multi-relational Attention Patterns. One additional advantage of using the multi-relational ontology is that we can relate the medical concepts in the EHR data to some auxiliary medical entities,

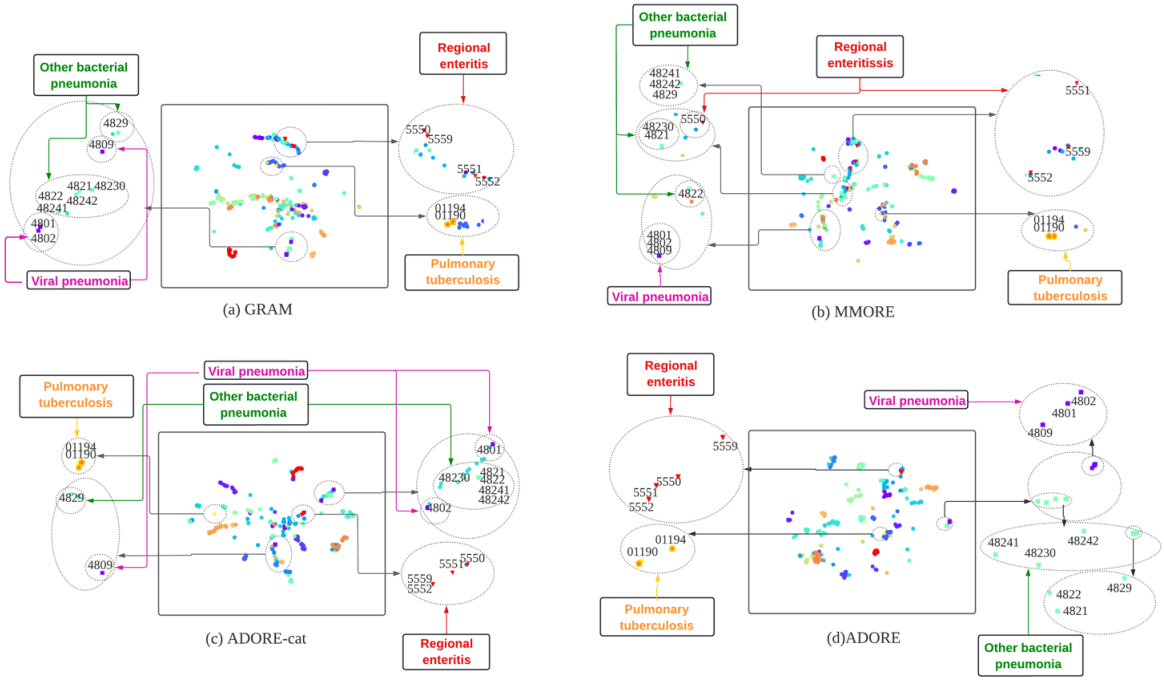


Fig. 6. Learned ontological embeddings for entities in Figure 5.

Table 6. Case studies of diagnoses with their associated relationships which are sorted by the attention learned from the model including "Abrasion or friction burn of face, neck, and scalp except eye, without mention of infection" (CCS code: 910.0), "Hypertensive heart and chronic kidney disease, unspecified, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified" (CCS code: 404.91), "Anemia in chronic kidney disease" (CCS code: 258.21).

CCS	Relationship	Incident entity	Attn
910.0	Morphology	Abrasion	0.7521
	Finding site	Skin of part of head and neck	0.1569
	Finding site	Skin structure of face	0.0911
404.91	Associated finding	Hypertensive disorder, systemic arterial	0.9965
	Has definitional manifestation	Blood pressure elevation	0.0032
	Finding site	Systemic arterial structure	0.0001
	Finding site	Cardiac structure	0.0001
258.21	Interprets	Hypertensive chronic kidney disease, unspecified, concentration	0.4684
	Clinical course	Chronic course-prolonged, duration	0.4056
	Finding site	Renal structure	0.1260

which may give rise to new insights into the data. Compared to the graph convolution neural networks, the proposed multi-relational attention layer in ADORE is able to identify the medical concepts which are crucial for specific prediction tasks as attributed by the corresponding relationships. Table 6 shows three case studies where the attention patterns derived using ADORE w.r.t the next-admission diagnosis prediction task are presented. For example, for the diagnosis *Abrasion or friction burn of face, neck, and scalp except for eye, without mention of infection* (CCS code: 910.0), three types of associated relationships are sorted by the attention values in the table. The **Morphology** relationship that links to *Abrasion* is found to be more important than the other two **Finding site** relationships. For the second case, the diagnosis *Hypertensive heart and chronic kidney disease, unspecified, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified* (CCS code: 404.91) follows a similar pattern. The **Associated finding** relationship with "Hypertensive disorder, systemic arterial" gains a higher attention value. Similarly, for the diagnosis *Anemia in chronic kidney disease* (CCS code: 258.21), the measurement of total hemoglobin (**Interprets** relationship) procedure describes the diagnosis better than the chronic course and renal structure relationships. These attention patterns present the types of features (expressed as associated relationships) of medical concepts which have more significant contribution to the prediction tasks.

6 CONCLUSION

In this paper, we propose ADORE which is a novel representation learning framework that can adaptively integrate both categorical and multi-relational medical ontologies with EHR data for medical concept embedding. First, it learns multiple basic embeddings for each category to avoid the potential inconsistency to be encountered between medical ontologies and EHR data, with the number of basic embeddings of each category automatically estimated. In addition, ADORE makes use of a category-aware attention network to allow more adaptive integration of the multi-relational ontology and the categorical ontology to preserve their distinct semantic relationships as far as possible. Empirical evaluation on real-world datasets demonstrates that ADORE can lead to improved predictive performance. Also, the learned medical concept representations are interpretable and can better align with the existing medical knowledge. For future work, we will explore methods to incorporate temporal information for the ontological embedding learning.

ACKNOWLEDGMENT

This research is partially supported by General Research Fund 12202117 and 12201219 from the Research Grants Council of Hong Kong.

REFERENCES

[1] Faisal Alshargi, Saeedeh Shekarpour, Tommaso Soru, A. Sheth, and Uwe Quasthoff. 2018. Metrics for Evaluating Quality of Embeddings for Ontological Concepts. In *Proceedings of the AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering (AAAI-MAKE 2019)*.

[2] Awais Ashfaq, Anita Sant'Anna, Markus Lingman, and Stawomir Nowaczyk. 2019. Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics* 97 (2019), 103256.

[3] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM International Conference on Management of Data*. 1247–1250.

[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*. 2787–2795.

[5] 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 International Conference on Knowledge Discovery and Data Mining* (2016), 1495–1504.

[6] 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 International Conference on Knowledge Discovery and Data Mining*. 787–795.

[7] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*. 3504–3512.
[8] Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. 2018. MiME: Multilevel medical embedding of electronic health Records for predictive healthcare. In *Advances in Neural Information Processing Systems*. 4552–4562.
[9] Youngduck Choi, Chill Yi-I Chiu, and David Sontag. 2016. Learning low-dimensional representations of medical concepts. In *AMIA Summits on Translational Science Proceedings*. 41.
[10] Trung Kien Dang, Xiang Lan, Jianshu Weng, and Mengling Feng. 2022. Federated Learning for Electronic Health Records. *ACM Transactions on Intelligent Systems and Technology* 13 (2022).
[11] Jason A. Fries, Ethan Steinberg, Saelig Khattar, Scott L. Fleming, Jose Posada, Alison Callahan, and Nigam H. Shah. 2021. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature Communication* 12 (2021), 104012.
[12] Shizhu He, Kang Liu, Guoliang Ji, and Jun Zhao. 2015. Learning to represent knowledge graphs with Gaussian embedding. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 623–632.
[13] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. 873–882.
[14] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge Graph Embedding via Dynamic Mapping Matrix.. In *Association for Computational Linguistics*. 687–696.
[15] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3 (2016), 160035.
[16] Igor Kulev, Pearl Pu, and Boi Faltings. 2018. A Bayesian Approach to Intervention-Based Clustering. *ACM Transactions on Intelligent Systems and Technology* (2018).
[17] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion.. In *Association for the Advancement of Artificial Intelligence*. 2181–2187.
[18] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM International Conference on Knowledge Discovery and Data Mining*. 1903–1911.
[19] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. KAME: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 743–752.
[20] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.
[21] 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*. 3111–3119.
[22] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.
[23] National Institutes of Health. [n. d.]. SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/index.html
[24] 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).
[25] X. Peng, G. Long, T. Shen, S. Wang, and J. Jiang. 2021. Sequential Diagnosis Prediction with Transformer and Ontological Representation. In *IEEE International Conference on Data Mining*. 489–498.
[26] Tom Pollard, Alistair Johnson, Jesse Raffa, Leo Celi, Roger Mark, and Omar Badawi. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data* 5 (09 2018), 180178. https://doi.org/10.1038/sdata.2018.178
[27] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web*. 593–607.
[28] Ying Sha and May D. Wang. 2017. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. 233–240.
[29] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of Graph Augmented Transformers for Medication Recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. 5953–5959.
[30] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2018. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics* 22, 5 (2018), 1589–1604.
[31] Lihong Song, Chin Wang Cheong, Kejing Yin, William K. Cheung, Benjamin C. M. Fung, and Jonathan Poon. 2019. Medical Concept Embedding with Multiple Ontological Representations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. 4613–4619.
[32] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes.. In *Association for the Advancement of Artificial Intelligence*. 1112–1119.

[33] WHO. 2004. ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision. https://apps.who.int/iris/handle/10665/42980

[34] WHOCC. [n. d.]. ATC. https://www.whocc.no/atc/

[35] Haoyi Xiong, Jinghe Zhang, Yu Huang, Kevin Leach, and Laura Barnes. 2017. Daehr: A Discriminant Analysis Framework for Electronic Health Record Data and an Application to Early Detection of Mental Health Disorders. *ACM Transactions on Intelligent Systems and Technology* 8 (2017).

[36] Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. *ArXiv Preprint ArXiv:1212.5701* (2012).

[37] Chunhong Zhang, Miao Zhou, Xiao Han, Zheng Hu, and Yang Ji. 2017. Knowledge Graph Embedding for Hyper-relational Data. *Tsinghua Science and Technology* 22, 02 (2017), 185–197.

[38] Muhan Zhang, Christopher R. King, Michael Avidan, and Yixin Chen. 2020. Hierarchical Attention Propagation for Healthcare Representation Learning. In *Proceedings of the 26th ACM International Conference on Knowledge Discovery & Data Mining*. 249–256.