Medical Concept Embedding with Multiple Ontological Representations

Lihong Song[1] , **Chin Wang Cheong**[1] , **Kejing Yin**[1] , **William K. Cheung** [1] , **Benjamin C. M. Fung**[2] and **Jonathan Poon**[3]

[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[2]School of Information Studies, McGill University, Montreal, Canada
[3]Hong Kong Hospital Authority, Hong Kong SAR, China
{lhsong, cwcheong, cskjyin, william}@comp.hkbu.edu.hk, ben.fung@mcgill.ca, jonathan@ha.org.hk

Abstract

Learning representations of medical concepts from the Electronic Health Record (EHR) has been shown effective for predictive analytics in healthcare. Incorporation of medical ontologies has also been explored to further enhance the accuracy and to ensure better alignment with the known medical knowledge. Most of the existing works assume that medical concepts under the same ontological category should share similar representations, which however does not always hold. In particular, the categorizations in medical ontologies were established with various factors being considered. Medical concepts even under the same ontological category may not follow similar occurrence patterns in the EHR data, leading to contradicting objectives for the representation learning. In this paper, we propose a deep learning model called MMORE which alleviates this conflicting objective issue by allowing multiple representations to be inferred for each ontological category via an attention mechanism. We apply MMORE to diagnosis prediction and our experimental results show that the representations obtained by MMORE can achieve better predictive accuracy and result in clinically meaningful sub-categorizations of the existing ontological categories.

1 Introduction

With the rapid growth of adoption of the Electronic Health Record (EHR), analyzing the EHR data to benefit the care for individual patients is attracting increasing attentions. Typically, the EHR data of a patient contains a set of clinical events including diagnoses, medications, procedures, laboratory tests, *etc.* Numerous efforts have been made to perform predictive analytics based on the medical concept representations learned from the EHR data, *e.g.* clinical event prediction [Choi *et al.*, 2017; Ma *et al.*, 2018], mortality prediction [Sha and Wang, 2017], among many others. Inspired by the well-adopted Word2Vce [Mikolov *et al.*, 2013b], these models usually learn a vector representation for each medical concept (*e.g.* a diagnosis code) from the co-occurrence information with the aim of making the frequently co-occurring medical concepts being close in the embedding space, such that the distance of two medical concepts in the embedding space can reflect their semantic closeness [Choi *et al.*, 2016a; Choi *et al.*, 2016c]. The medical concept representations being learned can then be further aggregated according to the patients' records and followed by the subsequent analytics, for instance the next-admission diagnosis prediction [Nguyen *et al.*, 2016; Ma *et al.*, 2017].

Despite the promising results obtained, one of the major challenges in this paradigm is that it relies on a large volume of training data which however is generally not easily available due to privacy concerns. Meanwhile, the representations learned purely from the (noisy) EHR data do not necessarily align with the existing medical knowledge, making the representations difficult to be interpreted by the clinicians. In view of these limitations, methods to inject existing medical knowledge into the representation learning process have been proposed recently [Choi *et al.*, 2017; Ma *et al.*, 2018]. Specifically, the medical ontologies which encode the relationships among the medical concepts in well structured formats (e.g. a knowledge graph or a hierarchical tree) can be utilized to guide the representation learning process.

The existing models incorporate the ontologies typically by assuming that the clinical concepts (*i.e.* the nodes in the ontologies) should be closer in the embedding space if they are closer in the ontologies (*e.g.* sharing the same category). This assumption is reasonable if the representations need only respect the relationships in the ontologies. However, it could hurt the accuracy of subsequent predictive analytics if we want the ontologies and the EHR data are to be respected at the same time to learn the representations. The main reason is due to the inherent inconsistency between the EHR co-occurrence and the ontologies. For example, 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" in the CCS (Clinical Classifications Software) ontology. However, they never co-occur in the EHR data as they are mutually exclusive. The benefit of introducing the ontologies and that of using the EHR data could easily "cancel out" each other, if the inconsistency issue is not carefully handled.

To alleviate this problem, we first relax the aforementioned assumption by allowing nodes close to one another in the ontologies not *necessarily* close in the embedding space. To achieve so, we borrow the idea of "multi-sense" from word embedding [Huang *et al.*, 2012] to allow each of the non-leaf nodes to carry multiple semantic meanings. To be more specific, we learn multiple vector-based representations, instead of only one, for the non-leaf nodes in the ontologies, and each of these representations is expected to correspond to a distinct "sense" and capture a particular group of lower level medical concepts with closer semantic meanings as reflected in the EHR data. Equipped with the attention mechanism [Bahdanau *et al.*, 2015] imposed on the ontologies, the desired separation becomes attainable. For instance, in the above example of diabetes mellitus, "T1DM" and "T2DM" would potentially be captured by two different "senses" of their common ancestor, as they barely co-occur in the EHR data and are generally very different in terms of pathology and therapy. Moreover, our proposed model, named ***M**edical concept embedding using **M**ultiple **O**ntological **RE**presentations (MMORE)*, also integrates the EHR co-occurrence statistics and the predictive task to produce more generalizable and interpretable representations. We further elaborate our model design with more technical details in Section 4.

We evaluated our proposed model using the open-source MIMIC-III dataset [Johnson *et al.*, 2016] and the results demonstrate that under our assumption and the novel strategy, the representations learned not only align better with the existing medical knowledge, but also achieve the desired separation as reflected in the EHR data, making it much more interpretable. Furthermore, the boost of predictive performance also validates the effectiveness of our strategy and the obtained representations. To the best of our knowledge, this is the first work attempts to address the inherent inconsistency between the EHR data and the medical ontologies when learning representations using both of them.

2 Related Works

The earliest works on medical predictive analysis mostly use one-hot representations, where the semantic relationships cannot be well preserved [Shickel *et al.*, 2018]. Recently, approaches similar to the Word2Vce method [Mikolov *et al.*, 2013b] are applied to learn the vector-based representations by considering the co-occurrence information of the clinicla concepts [Choi *et al.*, 2016a] or in the clinical narratives [Choi *et al.*, 2016c]. Some others make use of the sequential information in the EHR data. For example, RETAIN [Choi *et al.*, 2016b], Dipole [Ma *et al.*, 2017], MiME [Choi *et al.*, 2018] adopt the Recurrent Neural Networks to model the relationships among the medical concepts, guided by external predictive tasks in an end-to-end learning manner. However, these models rely on a large volume of data to effectively train their models.

Facing the problem of insufficient data, additional knowledge sources, *e.g.* medical ontologies, have been exploited to improve the quality of the learned representations and the predictive performance. For example the GRAM model [Choi *et al.*, 2017] develops the graph-based attention model to learn the representations from the knowledge graph. KAME model [Ma *et al.*, 2018] utilizes the ontology and the EHR data for learning concept representations. However, these models do not explicitly consider the inherent inconsistency between the EHR data and the ontologies, leaving learning effective representations from both EHR data and the ontologies an open question.

3 Notations and Preliminaries

3.1 Basic Notations

In this paper, we denote the EHR medical concepts, including diagnosis and medication codes, as $c_1, c_2, \ldots, c_{|C|} \in \mathcal{C}$ with the vocabulary size as $|\mathcal{C}|$. Each patient may have several hospital admissions, denoted as $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_t$, and each hospital admission consists of a subset of the medical concepts, *i.e.* $\mathcal{A}_t \subset \mathcal{C}$ for the t^{th} hospital admission. The co-occurrence statistics of the hospital admission $\mathcal{A}_t$ can therefore be represented by a binary vector $\mathbf{x}_t \in \{0,1\}^{|\mathcal{C}|}$ where the i^{th} entry of $\mathbf{x}_t$ equals to one if $c_i \in \mathcal{A}_t$, zero otherwise. The embedding matrices learned from the ontologies and the EHR co-occurrence statistics are denoted by $\mathbf{V}$ and $\mathbf{W}$ respectively.

3.2 Learning Representations from the Co-occurrence Statistics

The vector-based representations of the medical concepts can be learned using the approach similar to learning word embeddings [Mikolov *et al.*, 2013a]. Specifically, given a hospital admission $\mathcal{A}_t$ consisting of a set of l medical concepts $\{c_1, c_2 ... c_l\} \in \mathcal{C}$, we first take the average of their representations as the "context" of the admission, *i.e.*, $\mathbf{a}_t = (\sum_{k=1}^{l} \mathbf{w}_k)/l$. The basic idea of learning representations from the co-occurrence statistics is that the "context" should be able to predict the codes present in the hospital admission $\mathcal{A}_t$, which can be achieved by minimizing the negative log-probability of the codes being present in the admission conditioned on the "context", *i.e.*,

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

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

4 Proposed Model

In this paper, we aim to learn effective representations of the medical concepts from not only the EHR data, but also the medical ontologies by proposing the novel framework, Medical concept embedding with Multiple Ontological REpresentations (MMORE). Specifically, we consider ontologies that are represented in form of directed acyclic graphs (DAGs). Fig. 1 depicts the overall framework of our proposed model, where the matrices $\mathbf{W}$ and $\mathbf{V}$ in the center are the desired representations learned from the co-occurrence statistics and the ontologies separately. The upper right part of Fig. 1 corresponds to learning representations from EHR co-occurrence as described previously, and the lower right part is

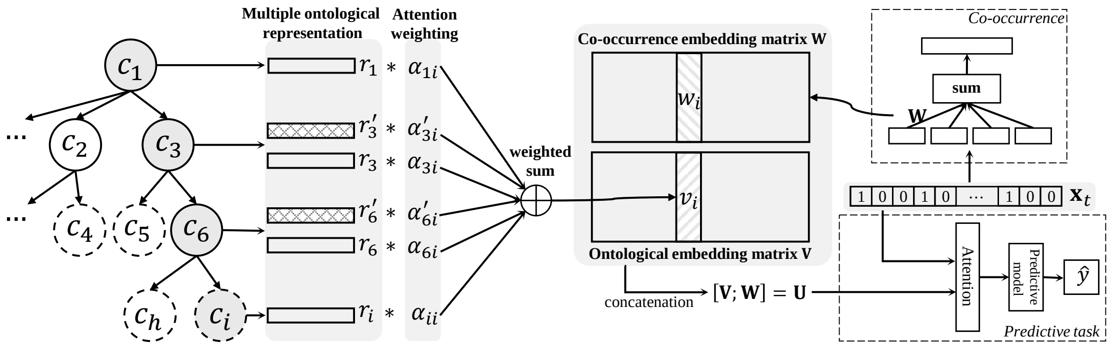


Figure 1: Framework overview of MMORE. The left part illustrates the idea of multiple ontological representation. The leaf nodes (dotted circles) are the medical concepts, while the non-leaf nodes (solid circles) represent the ontological categories. $\mathbf{r}$ denotes the basic embeddings of the nodes. MMORE learns multiple basic embeddings for the ancestors (except the root node) of the leaf nodes, *e.g.* $\mathbf{r}_3$ and $\mathbf{r}'_3$ for the node c_3. Then final ontological representation is derived by combining the basic embeddings via an attention mechanism. The upper right part illustrates learning the embeddings from the EHR co-occurrence and the lower right part is the predictive task.

the next-admission diagnosis prediction. The left part shows our approach of learning multiple ontological representations, which is detailed below.

4.1 Learning Multiple Ontological Representations

Given a medical ontology represented as a DAG with its leaf nodes corresponding to the set of all medical concepts $\mathcal{C}$, the "ancestor-descendant" relationships among the nodes of the DAG are used to learn the representations. Conventionally, each node in the DAG is embedded into a vector $\mathbf{r} \in \mathbb{R}^d$ as its basic embedding and the final representation of a node can then be derived from the convex combination of its all ancestors with the coefficients being inferred by the attention mechanism [Choi *et al.*, 2017]. However, as discussed earlier, this strategy will force the nodes to be close to its ancestors in the embedding space, leaving the inconsistency issue unresolved. To make the representations learned from the ontology more consistent with that from the EHR data, we propose to allow the non-leaf nodes (except the root node) in the DAG to carry multiple semantic meanings, or "senses". Formally, instead of learning one single vector representation for the non-leaf nodes (except the root node), we assign multiple basic embeddings, *e.g.* $\mathbf{r}, \mathbf{r}', \ldots$, to them. Without loss of generality, we assume two basic embeddings for each non-leaf node in this paper.[1] Following the GRAM model [Choi *et al.*, 2017], we use the attention mechanism to produce the final representations for the nodes in the DAG. The final representation $\mathbf{v}_i$ for the node c_i is given by:

$$\mathbf{v}_i = \sum_{j\ \in\ \text{ancestors}(i)} (\alpha_{ji}\mathbf{r}_j + \alpha'_{ji}\mathbf{r}'_j) + \alpha_{ii}\mathbf{r}_i, \tag{2}$$

where $\mathbf{r}_j$ and $\mathbf{r}'_j$ are the two basic embeddings for the node c_j, j denotes the index of a particular ancestor of the node c_i (its basic embedding is $\mathbf{r}_i$), and α denotes the attention weightings which are non-negative and sum up to one, *i.e.* $\sum_j(\alpha_{ji} + \alpha'_{ji}) + \alpha_{ii} = 1, \alpha_{ji} \geq 0\ \forall j$. Note that we fix $\alpha'_{1i} = 0$ for the root node. To compute the attention weightings, we first calculate the compatibility between the basic embeddings via a scoring function that is approximated by a single layer perceptron as:

$$f(\mathbf{r}_j, \mathbf{r}_i) = \mathbf{s}^\intercal \tanh\left(\mathbf{M}\begin{bmatrix}\mathbf{r}_j\\ \mathbf{r}_i\end{bmatrix} + \mathbf{b}\right), \tag{3}$$

where $\mathbf{s}$, $\mathbf{M}$ and $\mathbf{b}$ are the parameters to be learned. Then, the attention weightings can be obtained by applying the softmax function, *i.e.*,

$$\begin{aligned}\alpha_{ji} &= \frac{\exp(f(\mathbf{r}_j, \mathbf{r}_i))}{\omega_i}, \quad \alpha'_{ji} = \frac{\exp(f(\mathbf{r}'_j, \mathbf{r}_i))}{\omega_i},\\ \omega_i &= \sum_{k\in\text{ancestors}(i)} \left(\exp(f(\mathbf{r}_k, \mathbf{r}_i)) + \exp(f(\mathbf{r}'_k, \mathbf{r}_i))\right).\end{aligned} \tag{4}$$

In this paper, we focus on utilizing two medical ontologies, namely the "Clinical Classifications Software for ICD-9-CM"[2] (CCS) for the diagnosis codes and the "Anatomical Therapeutic Chemical classification system"[3] (ATC) for the medications. Both of them follow tree structures, where their non-leaf nodes are associated with a set of medical concept categories and nodes belonging to each category are regarded as clinically related.

4.2 Interpretability-Enhanced Predictive Analytics

We also incorporate the predictive task in our framework, which will "guide" the ontological representations to be learned. To be specific, we use the representations learned from the ontologies and the EHR co-occurrence to predict the next-admission diagnosis. Meanwhile, the attention mechanism is employed to further improve the interpretability and the prediction accuracy. The embedding matrices $\mathbf{W}$ and $\mathbf{V}$ are first row-wisely concatenated as the final representations of medical concepts, *i.e.*, $\mathbf{U} = [\mathbf{V}; \mathbf{W}]$. For simplifying the presentation, the formulations presented in this section are with respect to a single patient and we omit the subscript indexing the patient. Given a hospital admission $\mathcal{A}_t$ represented

[1] Generalizing the consideration to more than two basic embeddings will be studied in our future work.

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

[3] https://www.whocc.no/atc/

as a binary vector $\mathbf{x}_t$, we first calculate the intermediate representation for the hospital admission by retrieving the embeddings from $\mathbf{U}$ and summing them up, *i.e.* $\mathbf{a}_t = \mathbf{U}\mathbf{x}_t$. Then we compute the attention-based admission representation $\tilde{\mathbf{a}}_t$ as follows:

$$\tilde{\mathbf{a}} = \sum_{i=1}^{|\mathcal{A}|} \beta_i * \mathbf{u}_i, \quad \sum_{i=1}^{|\mathcal{A}|} \beta_i = 1, \ \beta_i \geq 0, \tag{5}$$

where we drop the subscript t denoting the t^{th} hospital admission for simplifying the notations, β denotes the attention weighting for the predictive task, and i denotes the index of the medical concepts contained in the admission $\mathcal{A}$. The attention weighting is computed by the softmax function:

$$\beta_i = \frac{\exp(g(\mathbf{a}, \mathbf{u}_i))}{\sum_{k=1}^{|\mathcal{A}|} \exp(g(\mathbf{a}, \mathbf{u}_k))}, \tag{6}$$

where the function $g(\cdot, \cdot)$ is approximated using a single layer perceptron with the same form as Eq. (3).

After the attention-based admission representation $\tilde{\mathbf{a}}_t$ being computed, it can be used as input to the prediction model. Without loss of generality, we use another single layer perceptron as the prediction model:

$$\hat{\mathbf{y}}_t = \text{softmax}\left(\tanh\left(\mathbf{Q}\tilde{\mathbf{a}}_t + \mathbf{k}\right)\right), \tag{7}$$

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 representation and $|N|$ is the number of classes. We compute the cross-entropy loss as the objective function for the predictive task as follows:

$$\mathcal{L}_p^{\text{pred}} = -\frac{1}{T-1} \sum_{t=1}^{T-1} \left[\mathbf{y}_t^\intercal \log(\hat{\mathbf{y}}_t) + (1-\mathbf{y}_t)^\intercal \log(1-\hat{\mathbf{y}}_t)\right], \tag{8}$$

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

By combining the objective functions of the predictive task and that of the EHR co-occurrence (Eq. 1), and taking average over all the patients, we can derive the overall objective function as follows:

$$\mathcal{L} = \frac{1}{N_p} \sum_{p=1}^{N_p} \left(\mathcal{L}_p^{\text{pred}} + \frac{1}{T_p} \sum_{t=1}^{T_p} \mathcal{L}_t^{\text{co-occur}} \right), \tag{9}$$

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

5 Experiments

We conduct experiments based on the MIMIC-III dataset to compare the performance of our proposed method MMORE with several state-of-the-art methods in terms of the prediction accuracy for the next-admission diagnosis prediction. Besides, we also evaluate the interpretability of the representations being learned using multiple ontological representations via case studies. In addition, we also present some particular phenotypes candidates derived from the attention weightings resulted from MMORE with related discussions regarding their quality.

Data	Model	20%	40%	60%	80%
Dx	RETAIN	0.4422	0.4447	0.4449	0.4545
	Med2Vec	0.5064	0.5187	0.5200	0.5290
	GRAM	0.4980	0.5218	0.5409	0.5498
	MMORE	**0.5205**	**0.5426**	**0.5548**	**0.5618**
Dx & Rx	RETAIN	0.4422	0.4447	0.4449	0.4547
	Med2Vec	0.4920	0.4967	0.4979	0.5110
	GRAM	0.5057	0.5285	0.5426	0.5548
	MMORE	**0.5243**	**0.5498**	**0.5619**	**0.5689**

Table 1: Accuracy@20 of diagnoses prediction, size of training data is varied (Dx is for diagnosis, and Rx is for medication)

Data set. MIMIC-III (Medical Information Mart for Intensive Care) [Johnson *et al.*, 2016] is a open-source dataset which comprises over 46k de-identified ICU patients collected over 11 years. In this paper, we focus on learning the representations of the diagnoses and medications.

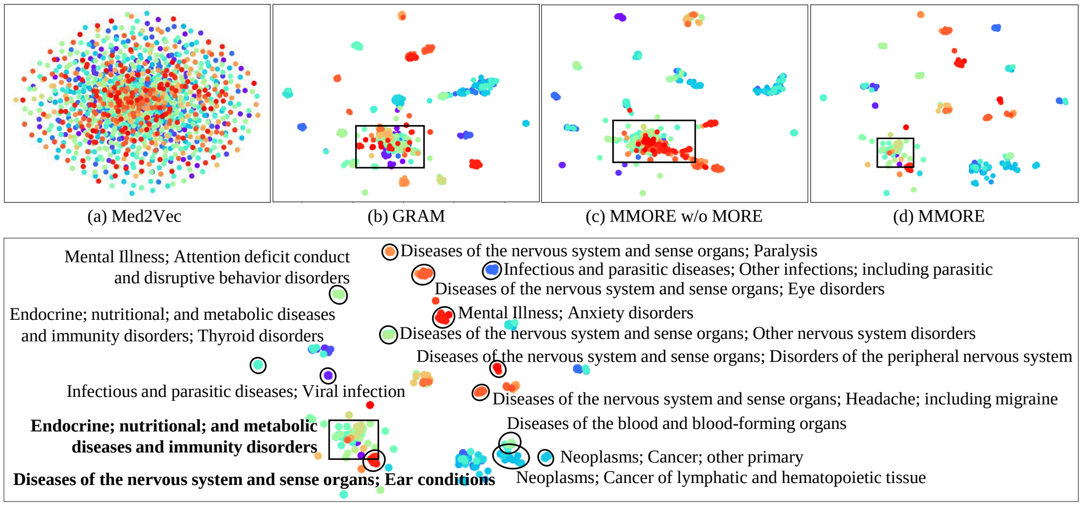
Data pre-processing. We extract the adult patients with at least two hospital admissions where diagnoses and medications are both present. We exclude the base type medications, *e.g.* D5W. In summary, we extract $5,404$ patients with average 2.6 hospital admissions per patient; the average numbers of diagnoses and medications in each admission are 12.3 and 41.1 respectively.

Baseline models. We compare the performance of our proposed framework against three state-of-the-art models:
RETAIN [Choi *et al.*, 2016b], which learns the medical concept representations and performs the heart failure prediction via the reversed RNN with the attention mechanism.
Med2Vec [Choi *et al.*, 2016a], which considers the medical concepts in consecutive admissions to capture their sequential and co-occurrence relationships.
GRAM [Choi *et al.*, 2017], which incorporates the medical ontology with an attention mechanism for the representation learning with the application to diagnosis prediction.

Experiment setup. We set the dimension of both the ontological embedding and the co-occurrence embedding to be 400 in our model. The embedding dimension of all baselines are set to be 800 for fair comparison as our model concatenates the two embedding matrices. The dimension of the hidden layer in the perceptrons used for the attention mechanism are set to be 100. The model is optimized using Adadelta [Zeiler, 2012] with batch size of 100.

5.1 Next-admission Diagnosis Prediction

We first evaluate the predictive performance by predicting the diagnoses in the next admission given the current one. In particular, We generate 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. We randomly split the data into 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. We measure the predictive



(a) Med2Vec (b) GRAM (c) MMORE w/o MORE (d) MMORE

(e) MMORE (with annotations)

Figure 2: Scatter diagrams of the representations learned from the CCS ontology. Nodes from 50 randomly selected lowest level categories in the CCS ontology are visualized. The upper row shows the visualization of the representation learned by different models. The bottom sub-figure is the enlarged version of sub-figure (d) with annotations. Each dot in the figure represents one diagnosis code, and the color represents the CCS category.

performance by *Accuracy@k*, which is defined as:

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

Results and discussion. The experimental results of the next-admission diagnosis prediction are summarized in Table 1. The results show that MMORE outperforms all the baselines, especially when the size of training data is small. This demonstrates that the superiority of our framework results from the explicit consideration of both the ontologies and the EHR co-occurrence, with the inherent inconsistency being well handled. Furthermore, we observe that the performance obtained by the models without using ontologies remain approximately the same (RETAIN) or even drop by up to 2.31% (Med2Vec) after adding the medications to the training data. The underlying reason may be that the next-admission diagnosis prediction is less sensitive to the medications, thus the relationships among medications cannot be well captured. By using the ontologies, GRAM and our MMORE model have marginal improvement when comparing the performance of using both diagnoses and medications for training and that of using only diagnoses. This implies that the ontologies could serve the role to "regularize" the learned representations of the medications. Overall, our proposed framework exhibits better predictive power, especially for the case of insufficient data.

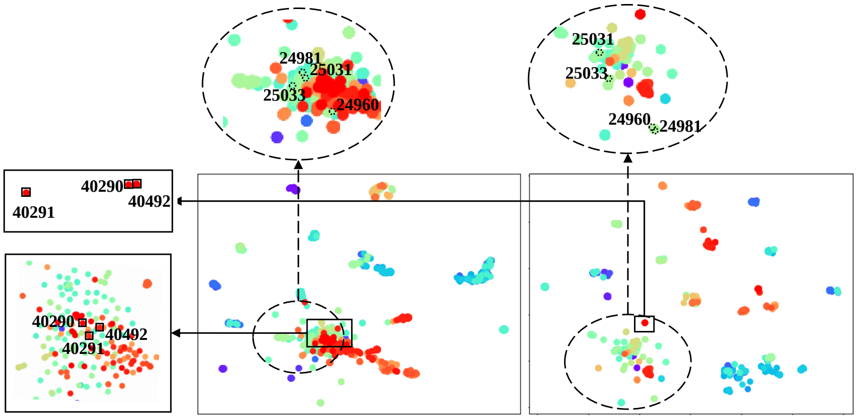
5.2 Interpretability of the Multiple Ontological Representation

To assess the interpretability of our multiple ontological representations, we use t-SNE [Maaten and Hinton, 2008] to visualize the representations learned from the medical ontologies. Due to space limit, we only exhibit the visualizations of the diagnoses. Specifically, we randomly select 50 categories from the third level counting from the bottom in the CCS ontology (excluding the leaf level), and visualize the representations of all the diagnoses (over 1,200 diagnoses in total) in the selected categories. For our proposed framework, we visualize the representations learned from the ontologies, *i.e.*, the columns of ontological embedding matrix $\mathbf{V}$. Note that for the Med2Vec and GRAM models, the representations being visualized are learned from EHR data and the ontology respectively.

The upper row of Fig. 2 are the representations learned by the different models, including Med2Vec (Fig. 2a), GRAM (Fig. 2b), MMORE without the Multiple Ontological REpresentations (MORE, *i.e.* learning only one basic embedding for each non-leaf node in the ontologies, Fig. 2c) and the MMORE framework (Fig. 2d). Each dot in the figure represents one diagnosis code, with its category indicated by the color of the dots. It is obvious from Fig. 2a that without using the ontology, the representations learned do not align with the existing medical knowledge. By adding the ontology information, GRAM model and our model without using the multiple ontological representations has much better alignment, yet the dots inside the rectangle in Fig 2b and Fig. 2c still do not form clear cluster structures that are consistent with the medical knowledge as indicated by the colors of the dots. The bottom figure (Fig. 2e) is an enlarged version of Fig. 2d with annotations. Evidently, the different categories shown in Fig. 2e are better separated, forming a clearer clustering structure comparing with the baseline models.

Case Study of the Interpretable Representations

To further demonstrate the effectiveness of introducing the multiple ontological representations, we conduct two case studies as visualized in Fig. 3. The first one is related to hypertensive heart diseases indicated by the solid rectangles where three diagnosis codes are identified (40291, 40290 and



(a) MMORE w/o MORE

(b) MMORE

24960	**Secondary diabetes mellitus** with neurological manifestations, not stated as uncontrolled, or unspecified
24981	**Secondary diabetes mellitus** with other specified manifestations, uncontrolled
25031	**Diabetes with other coma**, type I [juvenile type], not stated as uncontrolled
25033	**Diabetes with other coma**, type I [juvenile type], uncontrolled
40291	Unspecified hypertensive heart disease **with heart failure**
40290	Unspecified hypertensive heart disease **without heart failure**
40492	Hypertensive heart and chronic kidney disease, unspecified, **without heart failure** and with chronic kidney disease stage V or end stage renal disease

Figure 3: Case study of the learned ontological representations.

40492, see Fig. 3 for the annotation). The three diagnoses belong to the same lowest level category (the bottom level non-leaf node) in the CCS ontology ("hypertensive heart and/or renal disease"). Without using the multiple ontological representations, the three dots are close to each other (Fig. 3a), strictly following the information encoded in the CCS ontology. However, this is not desirable as the code 40291 (with heart failure) and 40290 (without heart failure) are exclusive thus will not co-occur in the EHR data. By adding the multiple ontological representations shown in Fig.3b, we observe that 40291 is separated from 40290, implying that the inconsistency issue between the medical ontology and the EHR data can be effectively alleviated by our proposed MMORE.

The second case, shown in the dotted circles in Fig. 3, relates to the two lowest level categories, namely "diabetes with neurological manifestations"(including diagnosis code 24960), and the other one is "diabetes with other manifestations" (including 24981, 25031 and 25033). In Fig. 3a, the code 24960 is far away from the other three codes, which follows the ontology structure. However, 24960 and 24981 are secondary diabetes while 25031 and 25033 are essential diabetes, which are two exclusive groups. With our multiple ontological representation framework, this relationship could be successfully captured as shown in Fig. 3b.

Phenotype 1 Dx: Atrial fibrillation; Congestive heart failure, NOS; ... *Rx: Warfarin; Heparin; ...*
Phenotype 2 Dx: Cirrhosis of liver w/o mention of alcohol; Dx: Alcoholic cirrhosis of liver; ... *Rx: Lactulose; Folic acid; ...*
Phenotype 3 Dx: Chronic airway obstruction, NEC; Dx: Obstructive chronic bronc w/ (acute) exacerbation; ... *Rx: Ipratropium bromide; Albuterol sulfate; ...*

Table 2: Three examples of derived phenotypes

5.3 Interpretation of the Predictive Attention Patterns (Phenotypes)

To further understand the attentions inferred from the predictive task, we apply the well-adopted dimensionality reduction tool, Non-negative Matrix Factorization (NMF) [Lee and Seung, 2001], to discover patterns from the learned attention weightings based on 4,000 patients that are randomly selected. The number of factors is set to be 15. The NMF model factorizes the input matrix into factors that group the related concepts together so that the patients can be better characterized by these factors. These factors typically can be called "phenotypes" and discovering phenotypes from EHR data has been regarded as a critical task in EHR data analytics [Kim *et al.*, 2017; Yin *et al.*, 2018]. Table 2 shows three examples of the inferred phenotypes, where the diagnoses (Dx) and the medications (Rx) in the first one are all related to heart disease, that in the second are all related to liver disease and that in the third are related with respiratory disease.

6 Conclusion

In this paper, we propose MMORE, a novel framework which leverages both medical ontologies and EHR data to learn robust and interpretable medical concept representations and meanwhile alleviates the inconsistency issue between the EHR data and the medical ontologies. Instead of learning one single basic embedding, MMORE tries to assign multiple basic embeddings to a single non-leaf node in the medical ontologies, which allows the final concept representations align better with the EHR data. The superiority of MMORE is empirically validated by the improvement of the predictive performance and shows better interpretability. For the future work, we will continuously focus on generalizing MMORE to facilitate more heterogeneous medical concepts and ontologies.

Acknowledgments

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

Table 3: Performance comparison in terms of accuracy@20 for next-admission diagnosis prediction, AUC for mortality prediction, and Mean Square Error (MSE) for length-of-stay prediction based on training sets of different sizes.

Task	Model	10%	20%	40%	80%
Next-admission diagnosis prediction	MLP	0.5017	0.5294	0.5509	0.5618
	Med2Vec	0.5085	0.5064	0.5187	0.5290
	GRAM	0.4850	0.5186	0.5382	0.5550
	MrMORE w/o Mr	0.5022	0.5378	0.5521	0.5628
	MrMORE-conv	0.4961	0.5286	0.5491	0.5598
	MrMORE	**0.5118**	**0.5410**	**0.5613**	**0.5690**
Mortality prediction	MLP	0.8792	0.9132	0.9246	0.9389
	Med2Vec	0.8375	0.8654	0.8794	0.9005
	MrMORE w/o Mr	**0.8874**	0.9117	0.9308	0.9440
	MrMORE-conv	0.8785	0.9044	0.9307	0.9431
	MrMORE	0.8844	**0.9140**	**0.9332**	**0.9496**
Length-of-stay prediction	MLP	1.012	0.8774	0.8684	0.7868
	Med2Vec	1.141	0.9769	0.9644	0.9223
	GRAM	1.112	1.059	1.015	0.9501
	MrMORE w/o Mr	1.109	0.8988	0.8350	0.8104
	MrMORE-conv	**1.008**	**0.8492**	0.8319	**0.7681**
	MrMORE	1.014	0.8678	**0.8312**	0.7820

Table 4: Case studies of diagnoses with their associated relationships which are sorted by attention learned from the model including "Abrasion or friction burn of face, neck, and scalp except eye, without mention of infection" (ICD-9 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" (ICD-9 code: 13.10), "Anemia in chronic kidney disease" (ICD-9 code: D63.1).

ICD	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 Chronic course-prolonged, duration	0.4056
	Finding site	Renal structure	0.1260

References

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[Choi *et al.*, 2016a] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504. ACM, 2016.

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

[Choi *et al.*, 2016c] Youngduck Choi, Chill Yi-I Chiu, and David Sontag. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41, 2016.

[Choi *et al.*, 2017] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. GRAM: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795. ACM, 2017.

[Choi *et al.*, 2018] Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. MiME: Multilevel medical embedding of electronic health records for predictive healthcare. In *Advances in Neural Information Processing Systems*, pages 4552–4562, 2018.

[Huang *et al.*, 2012] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 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*, pages 873–882. Association for Computational Linguistics, 2012.

[Johnson *et al.*, 2016] 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. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[Kim *et al.*, 2017] Yejin Kim, Robert El-Kareh, Jimeng Sun, Hwanjo Yu, and Xiaoqian Jiang. Discriminative and distinct phenotyping by constrained tensor factorization. *Scientific Reports*, 7(1):1114, 2017.

[Lee and Seung, 2001] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2001.

[Ma *et al.*, 2017] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1903–1911. ACM, 2017.

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

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[Mikolov *et al.*, 2013a] Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations*, pages 1–12, 01 2013.

[Mikolov *et al.*, 2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[Nguyen *et al.*, 2016] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. Deepr: a convolutional net for medical records. *IEEE Journal of Biomedical and Health Informatics*, 2016.

[Sha and Wang, 2017] Ying Sha and May D. Wang. 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*, pages 233–240. ACM, 2017.

[Shickel *et al.*, 2018] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 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):1589–1604, 2018.

[Yin *et al.*, 2018] Kejing Yin, William K. Cheung, Yang Liu, Benjamin CM Fung, and Jonathan Poon. Joint learning of phenotypes and diagnosis-medication correspondence via hidden interaction tensor factorization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3627–3633, 2018.

[Zeiler, 2012] Matthew D. Zeiler. ADADELTA: An adaptive learning rate method. *ArXiv Preprint ArXiv:1212.5701*, abs/1212.5701, 2012.