Enhancing Semantic and Structure Modeling of Diseases for Diagnosis Prediction

Hang Lv[1], Zehai Chen[1], Yacong Yang[1], Shuyao Pan, PhD[2], Bo Xiong, PhD[3], Yanchao Tan, PhD[1], Carl Yang, PhD[4]
[1] College of Computer and Data Science, Fuzhou University, Fuzhou, China
[2] Shengli Clinical Medical College, Fujian Medical University, Fuzhou, China
[3] Institute for Artificial Intelligence, University of Stuttgart, Stuttgart, Germany
[4] Department of Computer Science, Emory University, Atlanta, GA

Abstract

Electronic Health Records (EHRs) are valuable healthcare data, aiding researchers and doctors in improving diagnosis accuracy. Researchers have developed several predictive models by learning disease representations to forecast the potential diagnosis that patients may receive. However, existing studies usually ignore the fine-grained semantic and structure information in EHRs (e.g., the hierarchical relations between diseases and ICD-9 codes), which fails to provide accurate disease representation towards effective diagnosis prediction. To this end, we propose to enhance diagnosis prediction through LabCare, a framework with improved semantic and structure modeling of diseases in EHR data. LabCare can simultaneously capture rich semantic and structural relations among diseases and ICD-9 codes, which is achieved by innovatively integrating language models and box embeddings. Extensive experiments on two EHR datasets show that LabCare surpasses competitors, consistently achieving a 4.29% average improvement in Recall and NDCG metrics.

Introduction

Electronic Health Records (EHRs), serving as valuable healthcare data, provide researchers and doctors with the tools necessary to enhance the precision of diagnosis prediction-making[1,2,3]. These EHRs are repositories of extensive patient visit information, including diagnoses, medications, and procedures[4,5,6], laying the foundation for the development of more effective healthcare plans for patients.

The advent of machine learning techniques integrated with EHRs marks a significant milestone, heralding a new era of precise disease representation[1] [3,4]. This integration is crucial for diagnosis prediction, as it enables the accurate modeling of diseases, which can be further aggregated to represent patients, thereby facilitating effective and personalized diagnosis prediction. However, two challenges stand in the way of achieving accurate disease representation.

Challenge I: *How to effectively utilize semantic information to capture relations among diseases for modeling disease representation?* Although there have been efforts to utilize textual semantic information from EHRs to obtain disease representations[7,8], their efficacy in distinguishing relations among rare diseases remains constrained. Inspired by advancements in Language Models (LMs), some studies have attempted to apply these LMs for generating disease representations[9,10], where LMs have been pre-trained with the massive corpora and equipped with a broad spectrum of knowledge. However, due to the lack of specific clinical knowledge extracted from the medical corpus in training, it is difficult to model accurate relations between diseases via using LMs to comprehend diagnosis sets of disease names. For instance, while patients with hypertension may have an increased risk of developing coronary atherosclerosis, these two conditions are semantically distinct. This underscores the existing gap in the capacity of LMs to comprehend temporal relations between diseases. Moreover, LMs model disease similarity by prioritizing the textual semantics, thereby potentially failing to distinguish between diseases with similar names while in different categories. For instance, although "Nephritis nephropathy" and "Gouty nephropathy" have the phrase overlapping of "nephropathy", they belong to completely different categories (i.e., "Diseases of the Genitourinary System" and "Endocrine Diseases"). LMs may consider them as similar nephropathy and ignore the difference between the two diseases.

[1] We use the terms "representation" and "embedding" interchangeably in the remainder of this paper.

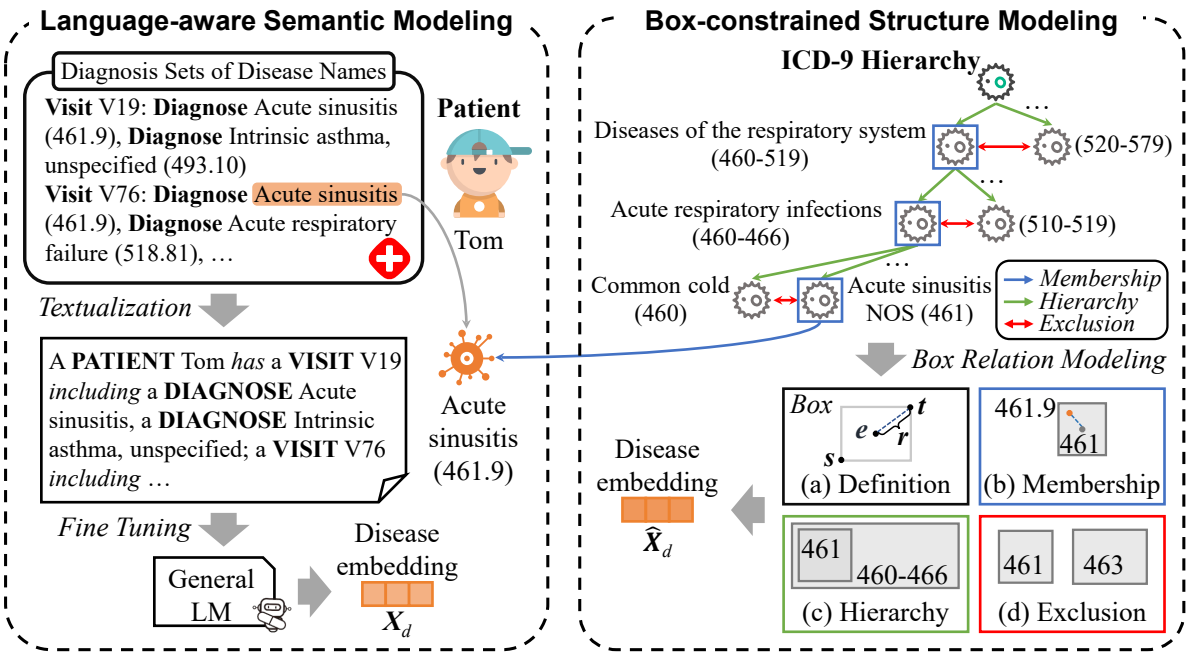


Figure 1: The overall framework of our proposed LabCare. Specifically, we use textualization to convert diagnosis sets of disease names from EHRs into composed text, and then fine-tune a General LM, thereby obtaining initial disease embeddings with rich semantics. Next, we model the structural relations among diseases and ICD-9 codes (i.e., membership, hierarchy, and exclusion) based on the ICD-9 Hierarchy using box embeddings. Finally, our LabCare performs diagnosis prediction via aggregating the learned disease embeddings.

Challenge II: *How to model and integrate the complex structure information into disease representation?* The abundant relations based on hierarchical structures of diseases, such as those provided by the Ninth Revision of International Classification of Diseases (ICD-9) codes in EHRs[2], further complicates the accurate representation of diseases. A common approach among these works involves capturing the hierarchical relations inherent in the ICD-9 Hierarchy, while tending to overlook the equally important membership and exclusive relations among diseases and ICD-9 code[2,11]. This neglect can lead to an inaccurate representation of disease and ICD-9 code relations. For instance, as depicted in Figure 1, if "Acute sinusitis (with ICD-9 code 461.9)" is categorized under ICD-9 code 461 (i.e., **membership** between 461.9 and 461), it logically follows that it falls within the broader category of ICD-9 code 460-466 (i.e., **hierarchy** between 461 and 460-466), and conversely, it should not be associated with ICD-9 code 460 (i.e., **exclusion** between 460 and 461). Based on the above relations, we can exactly distinguish "Nephritis nephropathy, unspecified (with ICD-9 code 583.9)" and "Gouty nephropathy, unspecified (with ICD-9 code 274.10)" which belong to different categories, thereby further capturing more accurate disease representation. However, how to simultaneously capture the three relations and integrate this complex structure information into disease representation remains unknown.

To address these challenges, in this work, we propose a novel framework to leverage LMs and box embeddings for enhancing disease representation and diagnosis prediction in healthcare data (named LabCare). Our method begins with a language-aware semantic modeling module. Specifically, to better understand semantic information from EHRs, we first convert patients' diagnosis sets of disease names with temporal structure into composed text via textualization. Then, we fine-tune a general LM based on the composed text and obtain initial disease embeddings with rich semantics. This module enables the model to better comprehend disease embeddings with clear relations in the order of visits, thereby distinguishing different diseases and laying the foundation for more accurate diagnosis predictions.

To further capture and integrate the complex structural relations from EHRs, we propose to project obtained initial

[2]https://www.cdc.gov/nchs/icd/icd9cm.htm

disease embeddings into points and ICD-9 code embeddings into boxes, where a box is an axis-aligned hyperrectangle with a geometric region. Two boxes can clearly "include" or "exclude" each other. In this way, we can utilize points and boxes to model the membership relations between diseases and ICD-9 codes. Meanwhile, we use boxes to capture the hierarchical and exclusive relations between ICD-9 codes. Upon box embeddings, we further obtain patient embeddings via aggregating the learned disease representations for diagnosis prediction, where disease representations integrate both semantics from diagnosis sets of disease names and structures from ICD-9 Hierarchy.

We conduct experiments on two real-world EHR datasets, a publicly accessible de-identified dataset named **MIMIC-III** and a multi-center ICU database named **eICU**, for diagnosis prediction, respectively. Extensive experimental results demonstrate that LabCare outperforms the state-of-the-art competitors, which constantly achieves an average of 4.29% improvement on Recall and NDCG metrics. Furthermore, insightful and interpretable case studies further demonstrate that our proposed LabCare can effectively capture semantic and structural relations among diseases and ICD-9 codes, thereby achieving accurate diagnosis prediction.

Related Work

In recent years, Language Models (LMs) have shown remarkable performance in Natural Language Processing (NLP) tasks[9,10,12], where LMs have been pre-trained with the extensive corpora and endowed with a wealth of general knowledge. This enables LMs to better comprehend semantic relations among diseases and model disease representations from the clinical text in EHRs, thereby supporting doctors and clinicians to make well-informed decisions[12,13,14,15]. However, most general LMs usually lack specialized clinical knowledge, which may cause them to overlook temporal relations among diseases and fail to distinguish different diseases.

More recently, some methods have attempted to leverage the structural relations between diseases and ICD-9 codes for precise disease representation learning[11,16]. However, most of them only consider the hierarchy, while ignoring other equally important relations, such as exclusion. Different from the above representation learning approaches, box embeddings offer a more natural and intuitive way to capture the complex structural relations[17,18,19,20]. Two boxes can naturally and clearly "include" or "exclude" each other, where a box is an axis-aligned hyperrectangle with a geometric region[21,22]. Therefore, box embeddings could aid in better understanding and modeling the structural relations among diseases and ICD-9 codes. Currently, no one has yet utilized LMs in combination with box embeddings to support and enhance the precision of diagnosis prediction-making. In the following section, we will introduce our LabCare framework and demonstrate how we leverage LMs and box embeddings to model accurate disease representations and achieve effective diagnosis predictions.

Method

In this section, we summarize the main modules of the LabCare framework in Figure 1 to provide an overview. LabCare performs diagnosis prediction through three modules. In the module of language-aware semantic modeling, we initialize disease embeddings via fine-tuning pre-trained Language Models (LMs) based on the obtained medical corpus set. Then, in the module of box-constrained structure modeling, we use box embeddings to model three relations, including membership, hierarchy, and exclusion, from the ICD-9 Hierarchy. After that, we further obtain patient embeddings via aggregating the learned disease representations for diagnosis prediction.

Language-aware Semantic Modeling. Drawing inspiration from the efficacy of representation learning via Language Models (LMs)[23,24,25,26], we propose to fine-tune pre-trained LMs based on medical corpus for the initialization of disease embeddings. As depicted in Figure 1, we first convert patients' diagnosis sets of disease names with temporal structure from EHRs into composed text through textualization. Then, we adopt a general LM DistilRoBERTa[3] as our starting point for fine-tuning[25], which can be flexibly replaced by other LMs (e.g., Clinical-BERT[27]). Next, we feed the text to the LM tokenizer and obtain the corresponding token list. The LM then learns the relations between these tokens to comprehend the textual information within patients' medical records, thereby yielding the fine-tuned LM. After obtaining the fine-tuned LM, we gain the initial disease embedding $\boldsymbol{X}_d$ based on the disease name, where

[3]https://github.com/huggingface/transformers/tree/main/examples/research_projects/distillation

the disease representation can capture both the temporal characteristics and clinical knowledge in the medical corpus. This improvement enables the model to better comprehend disease embeddings with clear relations in the order of visits, thereby distinguishing different diseases and laying the foundation for more accurate diagnosis predictions.

Box-constrained Structure Modeling. To begin with, we will introduce the box definition[17]. A box (hyperrectangle) can be described by two vectors (points). As shown in the lower right of Figure 1(a), we use the minimum point $\boldsymbol{s} \in \mathbb{R}^m$ and the maximum point $\boldsymbol{t} \in \mathbb{R}^m$ to represent a box $\boldsymbol{b}$, which is a m-dimensional hyperrectangle. The center point of the box is denoted as $\boldsymbol{e}$, and its radius is denoted as $\boldsymbol{r}$. The formula to calculate the volume of a box is given by $Vol(\boldsymbol{b}) = \prod_{k=1}^{m} (\boldsymbol{t}^k - \boldsymbol{s}^k)$, where k is the indicator of dimension. The intersection volume between two boxes $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$ is denoted as $Vol(\boldsymbol{b}_i \bigcap \boldsymbol{b}_j) = \prod_{k=1}^{m} \max(\boldsymbol{z}^k, 0)$, where $\boldsymbol{z}^k = \min(\boldsymbol{t}^k_{\boldsymbol{b}_i}, \boldsymbol{t}^k_{\boldsymbol{b}_j}) - \max(\boldsymbol{s}^k_{\boldsymbol{b}_i}, \boldsymbol{s}^k_{\boldsymbol{b}_j})$. Note that, the $\bigcap$ operator enables the calculation of conditional probability between two boxes, i.e., $p(\boldsymbol{b}_i|\boldsymbol{b}_j) = Vol(\boldsymbol{b}_i \bigcap \boldsymbol{b}_j)/Vol(\boldsymbol{b}_j)$.

Leveraging the advantages of box embeddings in modeling structural relations[17,22], we integrate three relations extracted from the ICD-9 Hierarchy to gain more accurate disease representations from both geometric and probabilistic perspectives. Specifically, we first project the initial disease embeddings $\boldsymbol{X}_d$ into points and ICD-9 code embeddings into boxes. Then, we use the geometric relations between a point and a box to model membership relations, and that between two boxes to model hierarchical and exclusive relations. Finally, we optimize these three relations via conditional probability. The right of Figure 1 shows how these relations are transformed into geometric constraints and we describe each relation in detail.

Membership Relation Since disease can have multiple ICD-9 codes (shown in Figure 1), we propose to leverage both points and boxes for modeling the membership relations among diseases and ICD-9 codes. Here, the center point of the box is denoted as $\boldsymbol{e}$, and its radius is denoted as $\boldsymbol{r}$. According to the property of mathematical definition on Membership (i.e., A point $\boldsymbol{d}_i$ is inside a box $\boldsymbol{b}_{c_j}\left(\boldsymbol{s}_{c_j}, \boldsymbol{t}_{c_j}\right)$ if and only if $\|\boldsymbol{d}_i - \boldsymbol{e}_{c_j}\| < \boldsymbol{r}_{c_j}$, which means $p\left(\boldsymbol{b}_{c_j}|\boldsymbol{d}_i\right) = 1 - tanh\left(\max\left\{0, \|\boldsymbol{d}_i - \boldsymbol{e}_{c_j}\| - \boldsymbol{r}_{c_j}\right\}\right) = 1.$), a disease d_i can be described by an ICD-9 code c_j. In this way, the corresponding geometric relation can be represented by making a point $\boldsymbol{d}_i \in \mathbb{R}^m$ being inside a box $\boldsymbol{b}_{c_j} \in \mathbb{R}^m$. Then, we define the membership objective function $\mathcal{L}_{Mem}$ by measuring the geometric membership (i.e., $\|\boldsymbol{d}_i - \boldsymbol{e}_{c_j}\| < \boldsymbol{r}_{c_j}$) as follows:

$$\mathcal{L}_{Mem}\left(\boldsymbol{b}_{c_j}, \boldsymbol{d}_i\right) = -\boldsymbol{b}_{c_j} \log\left(p\left(\boldsymbol{b}_{c_j}|\boldsymbol{d}_i\right)\right) - \left(1 - \boldsymbol{b}_{c_j}\right) \log\left(1 - p\left(\boldsymbol{b}_{c_j}|\boldsymbol{d}_i\right)\right), \tag{1}$$

where $\boldsymbol{d}_i$ denotes the disease embedding and $\boldsymbol{b}_{c_j}$ denotes the ICD-9 box embedding.

Hierarchical Relation Since a parent ICD-9 code can include its children geometrically (e.g., ICD-9 code 460-466 includes ICD-9 code 461), we propose to leverage the geometric insideness between the hyperrectangles of the corresponding box $\boldsymbol{b}(\boldsymbol{s}, \boldsymbol{t})$ for hierarchical relations. According to the property of mathematical definition on Hierarchy (i.e., A box $\boldsymbol{b}_{c_i}$ contains a box $\boldsymbol{b}_{c_j}$ if and only if $p\left(\boldsymbol{b}_{c_i}|\boldsymbol{b}_{c_j}\right) = 1.$), we transform the logical constraint into soft geometric constraint in the embedding space, where we propose a hierarchy loss as follows:

$$\mathcal{L}_{Hie}\left(\boldsymbol{b}_{c_i}, \boldsymbol{b}_{c_j}\right) = 1 - \frac{Vol\left(\boldsymbol{b}_{c_i} \bigcap \boldsymbol{b}_{c_j}\right)}{Vol\left(\boldsymbol{b}_{c_j}\right)}, \tag{2}$$

where box $\boldsymbol{b}_{c_i}\left(\boldsymbol{s}_{c_i},\ \boldsymbol{t}_{c_i}\right)$ contains box $\boldsymbol{b}_{c_j}\left(\boldsymbol{s}_{c_j},\ \boldsymbol{t}_{c_j}\right)$.

Exclusive Relation To properly model the exclusion between ICD-9 codes, we interpret the exclusion as geometric disjointness between boxes $\boldsymbol{b}(\boldsymbol{s}, \boldsymbol{t})$. According to the property of mathematical definition on Exclusion (i.e., a box $\boldsymbol{b}_{c_j}$ disconnects from a box $\boldsymbol{b}_{c_k}$ if and only if $p\left(\boldsymbol{b}_{c_j}|\boldsymbol{b}_{c_k}\right) = 0.$), we propose an exclusion loss $\mathcal{L}_{Ex}$ for exclusive relation modeling as follows:

$$\mathcal{L}_{Ex}(\boldsymbol{b}_{c_j}, \boldsymbol{b}_{c_k}) = \frac{Vol(\boldsymbol{b}_{c_j} \bigcap \boldsymbol{b}_{c_k})}{Vol(\boldsymbol{b}_{c_j}) \times Vol(\boldsymbol{b}_{c_k})}, \tag{3}$$

where box $\boldsymbol{b}_{c_j}(\boldsymbol{s}_{c_j},\ \boldsymbol{t}_{c_j})$ disjoints from box $\boldsymbol{b}_{c_k}(\boldsymbol{s}_{c_k},\ \boldsymbol{t}_{c_k})$.

By modeling the membership, hierarchical, and exclusive relations among diseases and ICD-9 codes, we effectively capture structure information about diseases. This enhancement equips our LabCare with a more sophisticated understanding of clinical knowledge.

Diagnosis Prediction. Based on the learned disease representations $\widehat{\boldsymbol{X}}_d$ via integrating LMs and box embeddings, we first adopt the self-attention mechanism[28] to obtain the historical visit embeddings of patient p_i. Following this, we generate patient embedding $\boldsymbol{p}_{i,T_i}$ via attention mechanism[29]:

$$\begin{aligned} \{w_{i,k}\}_{k=0}^{T_i-1} &= \text{softmax}\left(\text{MLP}\left(\{\text{Self-Att}\left(\boldsymbol{D}_{i,k}\right)\}_{k=0}^{T_i-1}\right)\right), \\ \boldsymbol{p}_{i,T_i} &= \sum_{k=0}^{T_i-1} w_{i,k}\boldsymbol{v}_{i,k}, \end{aligned} \tag{4}$$

where $\boldsymbol{D}_{i,k}$ is the set of disease embeddings diagnosed in the k-th visit of patient p_i, $\{\boldsymbol{v}_{i,k}\}_{k=0}^{T_i-1}$ denotes the historical visits of patient p_i, and $\{w_{i,k}\}_{k=0}^{T_i-1}$ denotes the weight assigned to each visit. Since the diagnosis prediction is a multi-label classification task, we use a dense layer with a softmax function to calculate the predicted probability:

$$\hat{\boldsymbol{v}}_{i,T_i} = \text{softmax}\left(\text{MLP}\left(\boldsymbol{p}_{i,T_i}\right)\right), \tag{5}$$

where $\hat{\boldsymbol{v}}_{i,T_i}$ is the prediction of p_i's T_i-th diagnosis. The objective function $\mathcal{L}_{Pred}$ for diagnosis prediction is listed:

$$\mathcal{L}_{Pred} = -\frac{1}{K}\sum_{i=1}^{K}\left(\boldsymbol{v}_{i,T_i}\log\left(\hat{\boldsymbol{v}}_{i,T_i}\right) + \left(1-\boldsymbol{v}_{i,T_i}\right)\log\left(1-\hat{\boldsymbol{v}}_{i,T_i}\right)\right), \tag{6}$$

where $\boldsymbol{v}_{i,T_i}$ is the ground-truth of p_i's T_i-th diagnosis and K is the number of patients. Finally, the overall objective function of LabCare is:

$$\mathcal{L} = \mathcal{L}_{Pred} + \lambda(\mathcal{L}_{Mem} + \mathcal{L}_{Hie} + \mathcal{L}_{Ex}), \tag{7}$$

where λ is a weight hyperparameter to control the regularization for box relation modeling.

LabCare combines the semantic modeling capabilities of LMs with the structure modeling features of box embeddings. This ensures that disease embeddings not only capture rich semantic information from EHRs but also incorporate structure information related to disease classification within the ICD-9 hierarchy. Finally, we achieve enhanced representations of diseases, leading to significantly improved performance in diagnosis prediction tasks.

Experimental Settings

Datasets. To verify the effectiveness of the compared methods, we use two real-world EHR datasets: **MIMIC-III**[30] and **eICU**[31]. The publicly available **MIMIC-III** dataset includes over forty thousand de-identified patients treated in critical care units at the Beth Israel Deaconess Medical Center from 2001 to 2010. We focus on **ICU admissions** within the dataset and utilize them for **diagnosis prediction**. Additionally, the **eICU** Collaborative Research Database is a multi-center ICU database monitoring over two hundred thousand ICU admissions across the United States. For the eICU dataset, we focus on **diagnostic informatics** and utilize it for **diagnosis prediction**. The sample characteristics of the two datasets are presented in existing works[30,31], and detailed statistics are shown in Table 1. Both datasets are split into training, validation, and test sets at a ratio of 7:1:2, with patients as the unit of segmentation.

Patient Cohort. For **MIMIC-III**, we extract a total of 7493 patients with at least two visits, considering their last visit as the diagnosis to be predicted and the previous visits as input to the model. For **eICU**, we extract patients with at least three visits, resulting in 23828 patients. In the visit records of these two datasets, diseases diagnosed during each visit are identified by ICD-9 codes[2,8], and the visit records are identified by unique patient IDs. For the hierarchy of ICD-9 codes, we obtain structure information from the website[4]. We sample all membership and hierarchical relations among diseases and ICD-9 codes, as well as exclusive relations between ICD-9 codes at the same level.

Evaluation Metrics. In assessing prediction performance, we employ two widely recognized metrics, Recall@k and NDCG@k. Recall@k measures the proportion of relevant diseases correctly predicted. NDCG@k (Normalized Discounted Cumulative Gain) simultaneously considers the correctness and ranking of each predicted disease. Consistent with ProCare[3], we set k to 5 and 10 to evaluate the predictive efficacy of our disease forecasting model.

[4]http://www.icd9data.com/

Dataset	# of patients	# of visits	Avg. visits per patient	# of unique ICD-9 codes	Avg. diagnosis codes per visit	Max diagnosis codes per visit
MIMIC-III	7,493	12,401	1.66	4,880	12.47	39.0
eICU	23,828	59,908	2.51	2,591	4.22	95.0

Table 1: Statistics of the datasets used in our experiments.

Methods for Comparison. We compare our proposed LabCare with the following baselines from two perspectives:

◇ **Interaction Modeling Methods.** These methods focus more on the interaction between diseases and visits to capture the structural characteristics of EHRs. Specifically, GRAM[32] leverages clinical knowledge graphs to learn representations of medical codes and predict visits using RNNs. KAME[33] focuses on predicting patients' future health information through a knowledge attention mechanism. MHM[1] models multi-modal clinical data-based hierarchical multi-label model for the diagnosis prediction task. TAdaNet[34] introduces a domain-knowledge graph and incorporates task-specific customization for diagnosis prediction. CGL[8] designs a collaborative graph learning model to explore patient-disease interactions and medical domain knowledge.

◇ **Dynamic Modeling Methods.** These methods focus more on the dynamic process of patients' status based on their visits in EHR data to model the temporal information. Specifically, RETAIN[35] utilizes gated recurrent units and attention mechanisms to predict patient diagnoses. Dipole[36] employs bidirectional RNNs and attention mechanisms to forecast patient visits. Timeline[37] formulates a time-aware disease progression function for predicting clinical events based on past visits. HiTANet[29] proposes a hierarchical time-aware attention network for risk prediction based on electronic health records. Chet[38] designs a context-aware dynamic graph learning mechanism to learn disease combinations and development. ProCare[3] enhances personalized diagnosis prediction by capturing disease severity, interaction, and progression in Electronic Health Records (EHR).

Implementation Details. We implement our model in PyTorch[5]. We employ the standard Adam optimizer with a learning rates of 1e-4. We set the embedding dimension dim to 128, the batch size to 128, and the weight hyperparameter λ to 3.0. For all baselines, we carefully tune their hyperparameters as suggested in the original papers to achieve their best performance.

Experiment Results

From Table 2, LabCare outperforms all baseline methods across four different evaluation metrics on both datasets. Compared to the second-best runner, LabCare achieves performance gains ranging from 2.15% in Recall@10 on eICU to 7.91% in NDCG@5 on MIMIC-III. This validates the effectiveness of our proposed semantic modeling and structure modeling for diagnosis prediction, as EHRs contain rich semantic information and structural characteristics.

Specifically, compared to ProCare, LabCare improves performance by 3.05% (achieved in Recall@5 on MIMIC-III) to 15.30% (achieved in NDCG@5 on eICU). This indicates that leveraging LMs for semantic modeling of diagnosis records contributes to performance enhancement, as LMs can incorporate prior knowledge from general corpora to learn the rich semantic information in EHRs. Compared to CGL, LabCare achieves performance gains ranging from 2.15% (achieved in Recall@10 on eICU) to 7.91% (achieved in NDCG@5 on MIMIC-III). This suggests that using box embeddings for structure modeling of ICD-9 codes is beneficial for performance improvement, where the three relations among diseases and ICD-9 codes in EHRs (i.e., membership, hierarchy, and exclusion) can be naturally and effectively captured and represented.

To better understand our proposed techniques, we conduct three ablation studies (shown in Table 2). Specifically, we respectively remove our general LM, fine-tuning, and box embedding techniques to obtain LabCare $_{l\text{-}}$, LabCare $_{f\text{-}}$, and LabCare $_{b\text{-}}$ models. Compared with LabCare $_{l\text{-}}$, LabCare learns semantic information carried by disease names in EHRs, resulting in performance gains ranging from 9.80% (achieved in Recall@10 on MIMIC-III) to 12.51% (achieved in NDCG@5 on MIMIC-III). Furthermore, the performance gains of LabCare over LabCare $_{f\text{-}}$ ranges

[5]https://pytorch.org/

Method	Recall@5	NDCG@5	Recall@10	NDCG@10	Recall@5	NDCG@5	Recall@10	NDCG@10
	MIMIC-III				eICU			
RETAIN	0.1510	0.4188	0.2134	0.3537	0.3213	0.3428	0.3901	0.3605
Dipole	0.1442	0.3999	0.2038	0.3378	0.3071	0.3274	0.3727	0.3452
GRAM	0.1429	0.4059	0.2112	0.3510	0.3049	0.3318	0.3862	0.3576
Timeline	0.1487	0.4123	0.2100	0.3482	0.3175	0.3376	0.3840	0.3548
KAME	0.1353	0.3992	0.2055	0.3070	0.2887	0.3268	0.3759	0.3126
MHM	0.1383	0.4080	0.2128	0.3481	0.2954	0.3340	0.3893	0.3547
TAdaNet	0.1433	0.4114	0.2172	0.3568	0.3056	0.3371	0.3972	0.3642
Chet	0.1457	0.3635	0.2051	0.3118	0.3104	0.2994	0.3722	0.3160
HiTANet	0.1502	0.4166	0.2122	0.3518	0.3204	0.3413	0.3881	0.3584
ProCare	0.1870	0.4383	0.2640	0.3823	0.4015	0.3622	0.4918	0.3922
CGL	0.1877	0.4676	0.2654	0.4012	0.4083	0.3958	0.5064	0.4279
LabCare $_{l\text{-}}$	0.1727	0.4485	0.2479	0.3888	0.3739	0.3731	0.4686	0.4043
LabCare $_{f\text{-}}$	0.1847	0.4886	0.2619	0.4154	0.4015	0.4014	0.4997	0.4310
LabCare $_{b\text{-}}$	0.1895	0.4821	0.2676	0.4137	0.4102	0.4004	0.5104	0.4299
LabCare	**0.1927**	**0.5046**	**0.2722**	**0.4301**	**0.4179**	**0.4176**	**0.5173**	**0.4449**

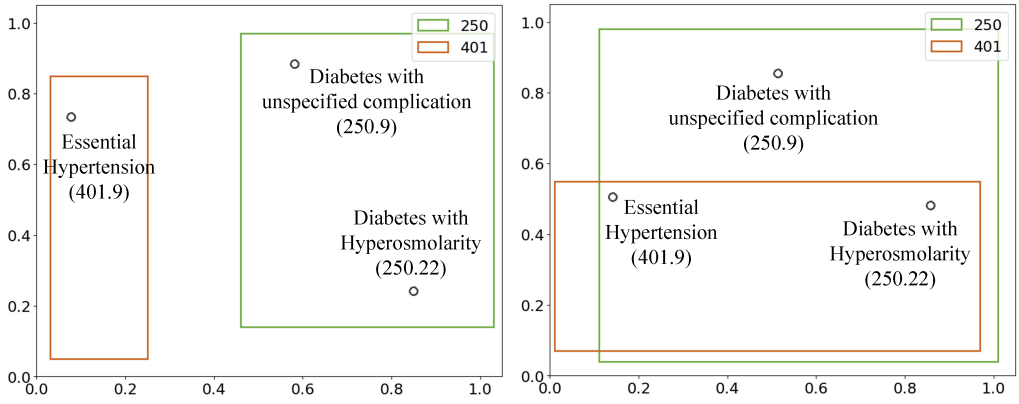
Table 2: Experimental results on two benchmark EHR datasets with Recall and NDCG. The best performances are highlighted in **boldface** and the second runners are underlined.

from 3.23% (achieved in NDCG@10 on eICU) to 4.33% (achieved in Recall@5 on MIMIC-III), where LabCare further comprehends the temporal information of patient visits and nuanced relations among diseases from EHRs via fine-tuned LM. The LabCare outperforms LabCare $_{b\text{-}}$ with gains ranging from 1.35% (achieved in Recall@10 on eICU) to 4.67% (achieved in NDCG@5 on MIMIC-III), mainly due to its additional modeling of the structural relations among diseases and ICD-9 codes.

Case Studies

		Jack's diagnoses		Mary's diagnoses
Ground-truth diagnoses of Visit 1-4	Visit 1	401.1 Benign essential hypertension 427.89 Other specified cardiac dysrhythmias . . .	Visit 1	274.9 Gout 584.9 Acute kidney failure 585.9 Chronic kidney disease . . .
	Visit 2	401.1 Benign essential hypertension 427.89 Other specified cardiac dysrhythmias . . .	Visit 2	274.00 Gouty arthropathy 585.9 Chronic kidney disease . . .
	Visit 3	401.9 Unspecified essential hypertension 427.89 Other specified cardiac dysrhythmias . . .	Visit 3	274.00 Gouty arthropathy 583.81 Nephritis and nephropathy 584.9 Acute kidney failure . . .
	Visit 4	401.9 Unspecified essential hypertension 414.01 Coronary atherosclerosis of native coronary artery 427.89 Other specified cardiac dysrhythmias . . .	Visit 4	274.01 Acute gouty arthropathy 584.9 Acute kidney failure . . .
Predictive diagnoses of Visit 4	LabCare$_{f\text{-}}$	401.9 Unspecified essential hypertension 427.89 Other specified cardiac dysrhythmias . . .	LabCare$_{b\text{-}}$	274.10 Gouty nephropathy 584.9 Acute kidney failure . . .
	LabCare	401.9 Unspecified essential hypertension 414.01 Coronary atherosclerosis of native coronary artery 427.89 Other specified cardiac dysrhythmias . . .	LabCare	274.01 Acute gouty arthropathy 584.9 Acute kidney failure . . .

Figure 2: Predictive diagnoses for patients Jack and Mary (pseudonyms) from the MIMIC-III dataset.



(a) Only Modeling Structural Relation (b) LabCare

Figure 3: Visualizations of syndromic diseases and their corresponding ICD-9 box embeddings.

Capturing temporal relations between diseases. As shown in Figure 2, for Jack's diagnoses with essential hypertension, LabCare forecasts the possibility of developing "Coronary atherosclerosis of the native coronary artery (with ICD-9 code 414.01)", which is consistent with the clinical knowledge (i.e., hypertension can increase the heart's workload on arteries, thereby leading to the risk of atherosclerosis). In contrast, LabCare $_{f\text{-}}$ merely predicts "Unspecified essential hypertension (with ICD-9 code 401.9)". This further underscores that the textualization and fine-tuning techniques in LabCare can effectively learn such temporal relations from semantic information in EHRs.

Differentiating diseases with similar names while in different categories. As shown in Figure 2, for Mary's diagnoses, LabCare distinguishes between "Acute gouty arthropathy (with ICD-9 code 274.01)" and "Gouty nephropathy (with ICD-9 code 274.10)", whereas LabCare $_{b\text{-}}$ fails to differentiate them. This confusion arises from the shared phrase, leading LabCare $_{b\text{-}}$ to misclassify them, while LabCare can differentiate them based on their distinct ICD-9 categories via box embeddings. This demonstrates the effectiveness of our box-constrained structure modeling in accurately distinguishing diseases with similar names while in different categories, which is a challenge for general LMs alone.

Capturing syndromic relations among diseases. As shown in Figure 3(b), despite the exclusion between ICD-9 codes 250 and 401, there is partial overlap between their corresponding box embeddings. This is mainly because Essential "Hypertension (with ICD-9 code 401.9)" and "Diabetes with Hyperosmolarity (with ICD-9 code 250.22)" are syndromic diseases, and patients often concurrently suffer from them in EHRs, leading to partial overlap between the boxes of ICD-9 codes. Compared to only model structural relations among diseases and ICD-9 codes (shown in Figure 3(a), LabCare can not only model both membership, hierarchy, and exclusion from the ICD-9 Hierarchy, but also capture the syndromic relations among diseases from patient-disease interactions in EHRs.

Conclusion

In this paper, we propose LabCare to achieve accurate diagnosis prediction, where we propose to enhance semantic and structure modeling of diseases for diagnosis prediction. Specifically, we simultaneously capture semantic and structural relations from diagnosis sets of disease names and ICD-9 hierarchy in EHRs, based on a fine-tuned language model and box embeddings. Subsequently, we perform accurate diagnosis prediction via aggregating the learned disease representation with rich semantic and structure information. Extensive experiments demonstrate the clear advantages of our LabCare over state-of-the-art baselines in diagnosis prediction, and insightful case studies show the accuracy and interpretability of our semantic and structural relation modeling.

For future works, it would be interesting to utilize more semantic information (e.g., ICD-9 names), explore more intricate structural relations among diseases and ICD-9 codes (e.g., intersection) for diagnosis prediction, and apply these insights to various crucial clinical tasks such as drug recommendation and early risk prediction.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants (No.6230071268) and the Natural Science Foundation of Zhejiang Province (LQ23F020007). Carl Yang was not supported by any fund from China.

References

1. Qiao Z, Zhang Z, Wu X, Ge S, Fan W. Mhm: Multi-modal clinical data based hierarchical multi-label diagnosis prediction. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2020. p. 1841-4.
2. Tan Y, Yang C, Wei X, Chen C, Liu W, Li L, et al. Metacare++: Meta-learning with hierarchical subtyping for cold-start diagnosis prediction in healthcare data. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2022. p. 449-59.
3. Tan Y, Zhou Z, Yu L, Liu W, Chen C, Ma G, et al. Enhancing Personalized Healthcare via Capturing Disease Severity, Interaction, and Progression. In: 2023 IEEE International Conference on Data Mining (ICDM). IEEE; 2023. p. 1349-54.
4. Xu R, Ali MK, Ho JC, Yang C. Hypergraph transformers for ehr-based clinical predictions. AMIA Summits on Translational Science Proceedings. 2023;2023:582.
5. Kim T, Heo J, Kim H, Shin K, Kim SW. VITA:'Carefully Chosen and Weighted Less' Is Better in Medication Recommendation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38; 2024. p. 8600-7.
6. Yu Z, Wang J, Luo W, Tse R, Pau G. MPRE: Multi-perspective Patient Representation Extractor for Disease Prediction. In: 2023 IEEE International Conference on Data Mining (ICDM). IEEE; 2023. p. 758-67.
7. Zhang J, Zhang X, Sun K, Yang X, Dai C, Guo Y. Unsupervised annotation of phenotypic abnormalities via semantic latent representations on electronic health records. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2019. p. 598-603.
8. Lu C, Reddy CK, Chakraborty P, Kleinberg S, Ning Y. Collaborative Graph Learning with Auxiliary Text for Temporal Event Prediction in Healthcare. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence; 2021. .
9. Yao Z, Cao Y, Yang Z, Deshpande V, Yu H. Extracting biomedical factual knowledge using pretrained language model and electronic health record context. In: AMIA Annual Symposium Proceedings. vol. 2022. American Medical Informatics Association; 2022. p. 1188.
10. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
11. Hansen ER, Sagi T, Hose K. Diagnosis Prediction over Patient Data using Hierarchical Medical Taxonomies. EDBT/ICDT Workshops. 2023.
12. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nature medicine. 2023;29(8):1930-40.
13. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. MedRxiv. 2023:2023-02.
14. Xu Y, Yang K, Zhang C, Zou P, Wang Z, Ding H, et al. VecoCare: visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23; 2023. p. 4921-9.
15. Kwon T, Ong KTi, Kang D, Moon S, Lee JR, Hwang D, et al. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38; 2024. p. 18417-25.
16. Wu Z, Xu D, Hu PJH, Huang TS. A hierarchical multilabel graph attention network method to predict the deterioration paths of chronic hepatitis B patients. Journal of the American Medical Informatics Association. 2023;30(5):846-58.
17. Onoe Y, Boratko M, Mccallum A, Durrett G. Modeling Fine-Grained Entity Types with Box Embeddings. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021. p. 2051-64.

18. Xiong B, Nayyeri M, Pan S, Staab S. Shrinking Embeddings for Hyper-Relational Knowledge Graphs. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics; 2023. p. 13306-20.
19. Zhang Y, Qin J, Feng C. PEB-TAXO: Projecting Entities as Boxes for Taxonomy Expansion. Neural Processing Letters. 2024;56(2):102.
20. Lv H, Chen Z, Yang Y, Ma G, Yanchao T, Yang C. BoxCare: A Box Embedding Model for Disease Representation and Diagnosis Prediction in Healthcare Data. In: Companion Proceedings of the ACM on Web Conference 2024; 2024. p. 1130-3.
21. Mei L, Mao J, Guo G, Wen JR. Learning probabilistic box embeddings for effective and efficient ranking. In: Proceedings of the ACM Web Conference 2022; 2022. p. 473-82.
22. Jiang S, Yao Q, Wang Q, Sun Y. A single vector is not enough: Taxonomy expansion via box embeddings. In: Proceedings of the ACM Web Conference 2023; 2023. p. 2467-76.
23. Wang L, Zhao W, Wei Z, Liu J. SimKGC: Simple Contrastive Knowledge Graph Completion with Pre-trained Language Models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022. p. 4281-94.
24. Yu D, Zhu C, Yang Y, Zeng M. Jaket: Joint pre-training of knowledge graph and language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36; 2022. p. 11630-8.
25. Tan Y, Zhou Z, Lv H, Liu W, Yang C. Walklm: A uniform language model fine-tuning framework for attributed graph embedding. Advances in Neural Information Processing Systems. 2024;36.
26. Ren X, Wei W, Xia L, Su L, Cheng S, Wang J, et al. Representation learning with large language models for recommendation. In: Proceedings of the ACM on Web Conference 2024; 2024. p. 3464-75.
27. Yan B, Pei M. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36; 2022. p. 2982-90.
28. Usama M, Ahmad B, Xiao W, Hossain MS, Muhammad G. Self-attention based recurrent convolutional neural network for disease prediction using healthcare data. Computer methods and programs in biomedicine. 2020;190:105191.
29. Luo J, Ye M, Xiao C, Ma F. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2020. p. 647-56.
30. Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3(1):1-9.
31. Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific data. 2018;5(1):1-13.
32. Choi E, Bahadori MT, Song L, Stewart WF, Sun J. GRAM: graph-based attention model for healthcare representation learning. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining; 2017. p. 787-95.
33. Ma F, You Q, Xiao H, Chitta R, Zhou J, Gao J. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In: Proceedings of the 27th ACM international conference on information and knowledge management; 2018. p. 743-52.
34. Suo Q, Chou J, Zhong W, Zhang A. Tadanet: Task-adaptive network for graph-enriched meta-learning. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2020. p. 1789-99.
35. Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. Advances in neural information processing systems. 2016;29.
36. Ma F, Chitta R, Zhou J, You Q, Sun T, Gao J. 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; 2017. p. 1903-11.
37. Bai T, Zhang S, Egleston BL, Vucetic S. Interpretable representation learning for healthcare via capturing disease progression through time. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining; 2018. p. 43-51.
38. Lu C, Han T, Ning Y. Context-aware health event prediction via transition functions on dynamic disease graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36; 2022. p. 4567-74.