Received August 9, 2020, accepted August 22, 2020, date of publication August 26, 2020, date of current version September 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019577

Patient Similarity via Joint Embeddings of Medical Knowledge Graph and Medical Entity Descriptions

ZHIHUANG LIN[1], **DAN YANG**[1], **AND XIAOCHUN YIN**[2]

[1]School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China
[2]Facility Horticulture Laboratory of Universities in Shandong, Weifang University of Science and Technology, Shouguang 262700, China

Corresponding author: Dan Yang (asyangdan@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 6167214, in part by the General Scientific Research Projects of Liaoning Province under Grant 2019LNJC07, and in part by the University of Science and Technology Liaoning Talent Project under Grant 601011507-22.

ABSTRACT With the prevalence and growing volume of Electronic Health Records (EHRs), there has been increasing interest in mining EHRs for improving clinical decision support. The accurate identification of patients with similar conditions based on EHRs is a key step in personalized healthcare. Existing studies model EHRs by medical knowledge graph embedding to learn the latent embeddings of medical entities (e.g., patients, medications, diagnoses and procedures). However, such precisely structured data is usually limited in quantity and in scope. Therefore, to enhance the quality of the embeddings it is important to consider more widely available medical information such as medical entity descriptions. In this paper we propose a novel framework, called Deep Patient Similarity (DeepPS). Specifically, DeepPS incorporates medical entity descriptions by augmenting the embeddings of medical entities and relations with the embeddings of words, which leverages both information from medical knowledge graph structures and the contexts of medical entity descriptions. Furthermore, DeepPS employs the embeddings to patient similarity learning by leveraging Siamese Convolutional Neural Network (CNN) with Spatial Pyramid Pooling (SPP). Extensive experiments on real datasets are conducted to show superior performance of our proposed framework.

INDEX TERMS Patient similarity, medical knowledge graph embedding, medical entity descriptions, Siamese CNN with SPP.

I. INTRODUCTION

Patient similarity learning [1] is a key and fundamental task in the medical healthcare domain, which aims to improve the doctors' diagnoses and the treatment of patients. With the tremendous growth of the adoption of EHRs, various sources of clinical information (e.g., demographics, diagnostic history, medications, procedures and laboratory test results) are becoming available about patients. This makes EHRs a valuable resource for identifying similar patients. The study of patient similarity aims at deriving a meaningful distance metric in the clinical field to measure the relative similarities among patients according to their health records. A proper similarity measure enables various downstream applications, such as personalized medicine [2], behavioral analysis [3] and medical diagnoses [4]–[11].

A. MOTIVATION

The precise patient similarity measures can group patients into cohorts effectively, which not only is beneficial to analyze the disease development trend of patients in similar cohorts, but also helps doctors make better clinical decisions to improve patients' health. For example, patients with higher risks of death during 48 to 72 hours of admission can be accurately identified, which enables doctors to take proactive measures to contain the risks. By understanding the development of the patients' conditions, doctors can design more targeted treatment plans for patients. This means that there will be fewer cases of over-treatment or under-treatment, patients will receive better advice, and patient care will become more personalized. Therefore, how to accurately and precisely measure patient similarity is an important and challenging issue.

With the increasing emergence of knowledge graphs, many world-leading researchers have successfully incorporated

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

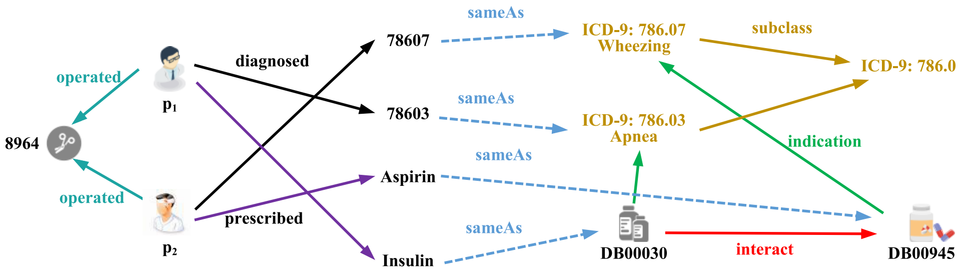


FIGURE 1. An annotated toy example of medical knowledge graph.

knowledge graphs into recommender systems [12]–[16] to improve the recommendation accuracy and explainability. In medical fields, it is of great significance to extract the valuable medical information from EHRs and build up a giant medical knowledge graph that reflects medical facts. Consequently, harnessing a well-built medical knowledge graph can provide more useful information for patient similarity learning. To efficiently exploit medical knowledge graphs in practice, many medical knowledge graph embedding approaches based on translation mechanism [17]–[19] and deep learning [20], [21] have been proposed to learn the embedding vectors of medical entities and relations. These approaches are demonstrated particular success in both performance and scalability, and are commonly adopted for deriving clinically meaningful representations of medical entities which are furthermore employed for patient profiling.

B. CHALLENGE

The key of patient similarity learning is to derive the effective representations of medical entities in the medical knowledge graph without loss of information. However, there are still significant challenges on learning effective vector representations of medical entities for deriving the patient similarity leveraging the medical knowledge graph:

1) Computational efficiency: Querying medical entities and relations based on conventional graph factorization algorithms have limitations in portability and scalability. The computational complexity becomes unfeasible when the medical graph reaches a very large scale.
2) Limited contents: Content limitations in EHRs constrain the researchers to establish more and more accurate relationships between medical entities of patients, which makes the final results less trustworthy for the complicated patients.
3) Exclusiveness: The embeddings are exclusive to entities/relations within the medical knowledge graph. As a result, the predictive ability is fundamentally limited by the information stored explicitly or implicitly in the medical knowledge graph, and the computation between the medical knowledge graph and medical entity descriptions cannot be handled.

C. SOLUTION

Taking into account all challenges mentioned above, we propose a novel patient similarity learning framework based on knowledge representation learning, which is able to take advantages of both medical knowledge graph and medical entity description. We name our framework as Deep Patient Similarity (DeepPS) throughout this paper. There are two parts in the proposed DeepPS: graph-text joint embedding and patient similarity learning. In graph-text joint embedding, we construct a medical knowledge graph (See Fig. 1.) from EHRs, ICD-9 ontology [22] and DrugBank [23], and obtain medical entity descriptions from Wikipedia[1] and EHRs. Then, DeepPS enables a joint embedding model to learn simultaneously from (1) medical knowledge triples that have been directly observed in a given medical knowledge graph, and (2) medical entity descriptions which have rich semantic information about these medical entities. More specifically, our joint embedding model consists of three parts: (1) **Medical Knowledge Graph Embedding**. We learn both medical entity and relation embeddings following the translation-based methods according to known triple facts in the medical knowledge graph. (2) **Medical Entity Description Embedding**. We learn the embedding representations of words in medical entity descriptions from word concurrences in text windows [24], where the distance between the words reflects the similarity between them. (3) **Alignment Model**. We learn to map both medical entity/relation and word embeddings into a unified low-dimensional semantic space. In the learning process, the joint embedding model encodes the semantics of medical entity descriptions to enhance the learning of medical knowledge graph embedding, and integrates such learned entity/relation embeddings to constraint their corresponding word embeddings in medical entity descriptions. In patient similarity learning, we propose a temporal patient representation based on the learned embeddings of medical entities. Afterwards, we deploy a patient similarity learning method based on the Siamese CNN model [25], [26] to compute the similarity score between all patient pairs.

D. CONTRIBUTIONS

The main distinctive technical contributions of our work are summarized as follows:

1) We develop the knowledge graph embedding model and medical entity description embedding model to learn medical entity/relation and word embeddings respectively, and leverage the alignment model to jointly map medical entity/relation and word

[1] https://encyclopedia.thefreedictionary.com/

embeddings into the same continuous vector space. The representations of medical entities enable the proposed framework to even effectively analyze the patient similarity.

2) We propose a novel method for modeling a patient based on the learned embeddings of medical entities and incorporate Siamese CNN with SPP as a deep learning model to measure the similarity between all patient pairs.

3) We conduct extensive experiments on large real datasets to show the efficiency of our proposed framework, which significantly outperforms the other baselines in terms of hospital readmission rate and incident rate difference for mortality. Moreover, comparative experiments are conducted between our proposed framework and the baselines in the performance of patient similarity learning, and the result of our experiments demonstrates that our proposed framework has the best performance.

The rest of this paper is organized as follows. Section 2 introduces the related work on patient similarity and medical knowledge graphs. We discuss our proposed framework in Section 3. The experimental results are reported in Section 4. Section 5 concludes the paper and our future work.

II. RELATED WORK

In this section, we first review some related work on evaluating the clinical patient similarity, and then review some relevant work associated with medical knowledge graphs.

A. PATIENT SIMILARITY

Recently, researchers have concentrated a lot of works on patient similarity measure in the field of health informatics. For example, Reference [27] deployed a cosine-similarity-based patient similarity metric (PSM) to weight the patient similarity measures. Reference [2] used the Tanimoto Coefficient (TC) to compute similarities between all patient pairs. Reference [28] proposed a locally supervised metric learning which is used for measuring similarities between patients represented by multi-dimensional time series. Nguyen *et al.* [29] proposed the sequential matching procedure to calculate the distance between two patients, which can utilize the sequential order of medical concepts. In addition, there are also a number of patient similarity measure methods taking into account the temporal information in EHRs. For example, Wang *et al.* [30] presented a convolutional matrix factorization for detection of temporal patterns, and Cheng *et al.* [1], [31] proposed an adjustable temporal fusion scheme using CNN extracted features. However, these methods are limited to patients with single disease, and the patients with a variety of diseases have been discarded directly. Therefore, Zhao *et al.* [32] designed a patient similarity label generation method for patients with multiple diseases, and converted the patient similarity measurement method into a multi-label classification problem.

B. MEDICAL KNOWLEDGE GRAPHS

Knowledge graphs have become ubiquitous nowadays as the backbone of multiple applications such as search engines and recommendation systems. Knowledge graphs provide an uncanny ability to capture the relationships between different entities by linking them through edges based on information extracted from various heterogeneous sources. Therefore, deploying knowledge graphs in the medical healthcare domain has proven to be an effective method to map relationships between the enormous variety and structure of healthcare data. An increasing number of knowledge graphs have been constructed from huge volumes of medical databases over the last years, such as Bio2RDF [33] and Chem2Bio2RDF [34]. However, there is little patients' clinical information within these medical knowledge graphs. STRIDE2RDF [35] and MCLSS2RDF [36] apply Linked Data Principles to represent electronic health records of patients. Unfortunately, such medical knowledge graphs are still limited to the interlinks from clinical data, which impedes its application in the medical healthcare domain.

III. THE PROPOSED FRAMEWORK

In this section, we first introduce the important notations used in this paper, and then explain our joint embedding model to construct the embeddings that jointly maps medical entities/relations and words of medical entity descriptions into the same continuous vector space. Finally, we present how to leverage Siamese CNN with SPP to measure the similarity between all patient pairs.

A. NOTATIONS

A patient's health record contains a sequence of medical concepts, which are recorded indicating the diagnoses, medications and procedures the patient suffered or received. The medical concepts are mapped to the International Classification of Disease (ICD-9) [37] and National Drug Code (NDC) [38]. We denote the set of all unique medical concepts from the EHR data as $\varepsilon = \{c_1, c_2, \ldots, c_{|\varepsilon|}\}$, where c_i is the medical concept and $|\varepsilon|$ is the number of unique medical concepts. A medical knowledge graph can be noted as $G = (E, R)$ which is a set of medical knowledge, where E is the set of medical entities including medical concepts in EHRs and medical entities in ICD-9 ontology and DrugBank, and R is the set of relations existing in the medical knowledge graph. Medical knowledge is comprised of entity-relation-entity triples in the form (e_h, r, e_t). Here $e_h \epsilon E$, $r \epsilon R$, and $e_t \epsilon E$ denote the head, relation and tail of a medical knowledge triple, respectively. For instance, in Fig. 1 a triple $(p_1, \mathit{prescribed}, \mathit{Insulin})$ indicates that there is a relationship *prescribed* from the patient p_1 to the medication *Insulin*. And a triple $(\mathit{Insulin}, \mathit{sameAs}, \mathit{DB}00030)$ indicates the medical concept *Insulin* from EHRs has the *sameAs* relation with the medical entity *DB*00030 from DrugBank. Given a medical entity e, we let $\mathit{text}(e) = w_1, w_2, \ldots, w_n$ be the sequence of words associated with medical entity e. In other words, $\mathit{text}(e)$ is the description of medical entity e (See Fig. 2.).

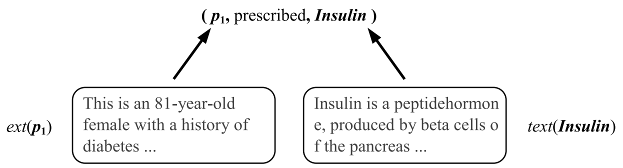

(p₁, prescribed, Insulin)
ext(p₁)
This is an 81-year-old female with a history of diabetes ...
Insulin is a peptidehormon e, produced by beta cells o f the pancreas ...
text(Insulin)

FIGURE 2. Example of medical entity descriptions.

We try to learn the embeddings **e**, **r** and **w** for each medical entity e, relation r and word w respectively. The descriptions of all medical entities are used as a text corpus Δ and the vocabulary of words from text corpus Δ is V. The union vocabulary of medical entities and words is $\Omega = E \cup V$.

B. JOINT EMBEDDING MODEL COMBINING MEDICAL KNOWLEDGE GRAPH AND MEDICAL ENTITY DESCRIPTIONS

Our joint embedding model consists of three parts: medical knowledge graph embedding, medical entity description embedding and alignment model. Fig. 3 demonstrates the overall architecture of our joint embedding model. In the following section, we introduce the three parts in details.

1) MEDICAL KNOWLEDGE GRAPH EMBEDDING

A medical knowledge graph $G = (E, R)$ consists of a set of interconnected medical entities and their relations, where medical entities E and relations R can be different types. Given a medical fact triple $(e_h, r, e_t)\epsilon G$, the letters $\mathbf{e}_h$, $\mathbf{r}$, $\mathbf{e}_t$ are characterized as the corresponding embedding representations of e_h, r, e_t. Recently, significant advancement has been made in using the translation-based method to train medical knowledge graph embedding. To characterize a medical fact triple (e_h, r, e_t), the translation-based models follow a common assumption $\mathbf{e}_h^* + \mathbf{r} \approx \mathbf{e}_t^*$, where $\mathbf{e}_h^*$ and $\mathbf{e}_t^*$ are either the embedding representations of e_h and e_t, or the transformed vectors under a certain transformation w.r.t. relation r. TransR [19] is a state-of-the-art translation-based embedding approach. It has achieved promising results in knowledge graph completion and link prediction from text.

Consider the above reason, for each medical fact triple (e_h, r, e_t), medical entities embeddings are set as $\mathbf{e}_h$, $\mathbf{e}_t \epsilon \mathbb{R}^k$ and relations embeddings are set as $\mathbf{r} \epsilon \mathbb{R}^d$. For each relation r, we set a projection matrix $\mathbf{H}_r \epsilon \mathbb{R}^{k\times d}$, which may project medical entities from entity space to relation space. We define the translations between medical entities and get the energy function $z(e_h, r, e_t)$ as:

$$z(\mathbf{e}_h, \mathbf{r}, \mathbf{e}_t) = b - \|\mathbf{e}_h\mathbf{H}_r + \mathbf{r} - \mathbf{e}_t\mathbf{H}_r\|_{L1/L2} \tag{1}$$

where b is a constant for bias designated for adjusting the scale for better numerical stability.

Then, we define the following conditional probability of a medical fact triple (e_h, r, e_t) in a medical knowledge graph:

$$P(e_h|r, e_t) = \frac{exp\{z(\mathbf{e}_h, \mathbf{r}, \mathbf{e}_t)\}}{\sum\limits_{\hat{e}_h \in \Omega} exp\{z(\hat{\mathbf{e}}_h, \mathbf{r}, \mathbf{e}_t)\}} \tag{2}$$

and $P(e_t|e_h, r)$, $P(r|e_h, e_t)$ can be defined in the same way by choosing corresponding normalization terms respectively. We define the likelihood of observing a medical fact triple (e_h, r, e_t) as:

$$\begin{aligned} \pounds(e_h, r, e_t) = logP(e_h|r, e_t) + logP(e_t|e_h, r)& \\ + logP(r|e_h, e_t)& \end{aligned} \tag{3}$$

The goal of medical knowledge graph embedding is to maximize the conditional likelihoods of existing fact triplets in the medical knowledge graph. Based on (3), the objective function of medical knowledge graph $G = (E, R)$ can be defined as follows:

$$\pounds_G = \sum_{(e_h, r, e_t)\in G} \pounds(e_h, r, e_t) \tag{4}$$

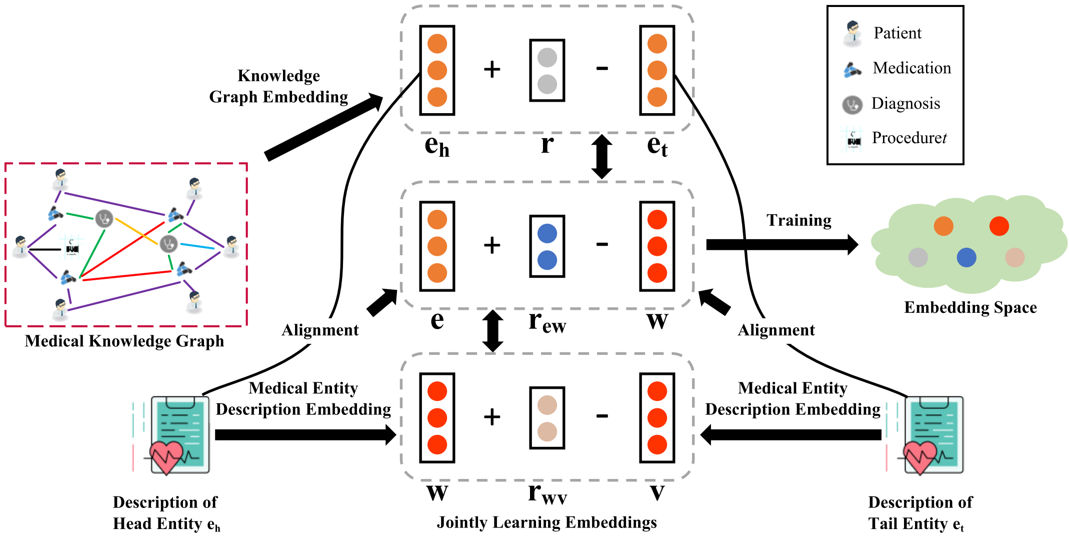


FIGURE 3. Overall architecture of our joint embedding model.

2) MEDICAL ENTITY DESCRIPTION EMBEDDING

The medical entity descriptions contain rich and important knowledge information, and it is also one of the multi-source information that can interact with the medical knowledge bases (e.g., MIMIC-III, DrugBank and ICD-9 ontology). It is assumed that the medical entities are similar if they have similar entity descriptions. To encode the rich semantic representations of words in given medical entity descriptions, we adopt the text model proposed in [39] as our medical entity description embedding model. It learns word embeddings by capturing the co-occurrence of words observed in a text corpus Δ. In the medical entity description embedding model, we assume any pair of words w and v that concur in some fixed-size context windows are of certain r_{wv} but r_{wv} is a hidden variable, and the goal is to fit the concurring pairs of words. Therefore, the energy function $z(w, r_{wv}, v)$ evaluating the co-occurrence of two words w and v based on their embeddings is defined as follows:

$$z(\mathbf{w}, \mathbf{r}_{wv}, \mathbf{v}) = b - \|\mathbf{w} + \mathbf{r}_{wv} - \mathbf{v}\|_{L1/L2} \tag{5}$$

where $\mathbf{w}$ and $\mathbf{v}$ are the embedding representations of two words w and v respectively.

Then, the conditional probability of a target word w appearing close to a context word v (within a context window of a certain length) can be defined as follows:

$$P(w|r_{wv}, v) \triangleq P(w|v) = \frac{exp\{z(\mathbf{w}, \mathbf{r}_{wv}, \mathbf{v})\}}{\sum\limits_{\hat{v}\in V} exp\{z(\mathbf{w}, \mathbf{r}_{w\hat{v}}, \hat{\mathbf{v}})\}} \tag{6}$$

Subsequently, the objective function of the medical entity description embedding model is to maximize the likelihood of the concurrences of pairs of words in text windows:

$$£_W = \sum_{(w,v)\in C} \#(w, v) logP(w|v) \tag{7}$$

where C is all the distinct pairs of words concurring in text windows of a fixed size, and $\#(w, v)$ is the number of times (w, v) appears in the text corpus Δ.

3) ALIGNMENT MODEL

The embedding vectors of medical entities/relations from knowledge graph embedding model and word embeddings from medical entity description embedding model do not interact, and they can be placed in different subspaces of the vector space. To address this issue, based on medical knowledge graph embedding and medical entity description embedding, we introduce the alignment model to jointly embed medical entity/relation and word embeddings into the same vector space.

Inspired by the translation-based methods for knowledge representation learning such as TransE [17], it is straightforward to regard the alignment as a special relation between the medical entity and each word in its description, and perform an alignment-specific translation operation between the medical entity and each word in its description to learn joint embeddings.

Formally, given the medical entity e and each word w in its description, we assume there is an alignment relation r_{ew} so that $\mathbf{e} + \mathbf{r}_{ew} \approx \mathbf{w}$. The energy function of joint embeddings is thus defined as:

$$z(\mathbf{e}, \mathbf{w}) = b - \|\mathbf{e} + \mathbf{r}_{ew} - \mathbf{w}\|_{L1/L2} \tag{8}$$

The conditional probability $P(w|e)$ of predicting w given e can be defined as follows:

$$P(w|e) = \frac{exp\{z(\mathbf{e}, \mathbf{w})\}}{\sum\limits_{\hat{w}\in V} exp\{z(\mathbf{e}, \hat{\mathbf{w}})\}} \tag{9}$$

We also define $P(e|w)$ in the same way by revising the normalization term:

$$P(e|w) = \frac{exp\{z(\mathbf{e}, \mathbf{w})\}}{\sum\limits_{\hat{e}\in E} exp\{z(\hat{\mathbf{e}}, \mathbf{w})\}} \tag{10}$$

Then the objective function of alignment model can be defined as follows:

$$£_A = \sum_{e\in E} \sum_{w\in text(e)} [\log P(w|e) + \log P(e|w)] \tag{11}$$

4) OPTIMATION AND TRAINING

To jointly learn the embeddings of words and medical entities/relations by simultaneously maximizing the sum of the three logarithm likelihood of objective functions just as follows:

$$£(T) = £_G + £_W + £_A + \gamma\Theta(T) \tag{12}$$

where T stands for the embeddings of medical entities, relations and words, γ is a hyper-parameter weighting the regularization factor $\Theta(T)$, which is defined as follows:

$$\Theta(T) = \sum_{e\in E} [\|e\| - 1]_+ + \sum_{r\in R} [\|r\| - 1]_+ + \sum_{w\in V} [\|w\| - 1]_+ \tag{13}$$

where $[x]_+ = max(0, x)$ denotes the positive part of x. The regularization factor will normalize the embeddings during learning. And we adopt stochastic gradient descent (SGD) [40] to maximize the transformed objective function.

Optimizing objective functions (4) and (7) in (12) are computationally expensive, as calculating them need to sum over the entire set of medical entities, relations and words. To address this problem, we use negative sampling (NEG) to transform the original objective, i.e., Equation (12) to a simple objective of the binary classification problem—differentiating the observed data from noise.

For (4), we should transform $log\ P(e_t|e_h, r)$, $log\ P(e_h|r, e_t)$ in (3). Taking as $P(e_h|r, e_t)$ an example, we maximize the following objective function instead of it:

$$log\ \sigma(z(\mathbf{e}_h, \mathbf{r}, \mathbf{e}_t)) + \sum_{i=1}^{\mu} \mathbb{E}_{\tilde{e}_h^i \sim z_{neg}(\{\tilde{e}_h, r, e_t)\}}[\sigma(z(\tilde{\boldsymbol{e}}_h^i, \mathbf{r}, \mathbf{e}_t))] \tag{14}$$

where μ is the number negative examples to be discriminated for each positive example, $\sigma(x) = 1/(1 + exp(-x))$ is the sigmoid function. $\{(\tilde{e}_h, r, e_t)\}$ is the invalid triple set, and z_{neg} is a function randomly sampling instances from $\{(\tilde{e}_h, r, e_t)\}$. When a positive triple $(e_h, r, e_t)\epsilon G$ is selected, to maximize (14), μ negative triples are constructed by sampling medical entities from an uniform distribution over E and replacing the head of (e_h, r, e_t). The transformed objective of $log\ P(r|e_h, e_t)$, $log\ P(e_t|e_h, r)$ are maximizing in the same manner, but for $log\ P(r|e_h, e_t)$, the negative relations are sampled from a uniform distribution over R to corrupt the positive relation $r\epsilon\ (e_h, r, e_t)$. We iteratively select random mini-batch from the training set to learn embeddings until converge. Thus, we can also simplify (7) and maximize it in the same way by choose the corresponding negative distribution.

In our training process, we implement a multi-threading version to learn the representations for a better efficiency. To avoid overfitting, we initialize entity and relation embeddings with results of TransE, and initialize projection matrix as identity matrix.

C. PATIENT SIMILARITY LEARNING

Inspired by the text similarity problem tackled by the Siamese LSTM model [41], it is available to measure the similarity between all patient pairs using Siamese CNN. Each of the twin subnetworks of Siamese CNN uses this same CNN architecture. However, there is a technical issue in the training and testing of CNN: the fixed-size patient representations are taken as the input of CNN, which limits both the aspect ratio and the scale of the input. When applied to the patient representations of arbitrary sizes, current methods mostly fit the input to the fixed size, either via cropping [42] or via warping [43]. But the cropped region may not contain the entire object, while the warped content may result in deformation. Therefore, in order to remove the fixed-size constraint of the CNN, we add an SPP layer [44] on top of the last convolutional layer. The architecture of CNN is adapted by introducing the SPP layer. The new architecture of CNN is called SPP-net. Specifically, we utilize the learned embeddings to construct the temporal patient representation, and then train a Siamese CNN model with SPP, which first maps the pair of temporal patient representations with different dimensions to the fixed-size vectors respectively and then we use the Euclidean distance as the negative similarity function to express the degree of relatedness between the pair of patients. That is, the Euclidean distance between the two patient vectors is taken as the final similarity score.

1) TEMPORAL PATIENT REPRESENTATION

The process of patient similarity learning involves the construction of patient representations based on the medical entity embeddings we have learned. A straightforward representation of a patient is to convert all medical concepts in his medical history to medical concepts vectors, and then summing all those vectors to obtain a single representation vector. However, this kind of patient representation ignores the temporal information of medical concepts from EHRs. Therefore, we adopt a temporal patient representation method based on the happening timestamps of medical concepts in an increasing order, i.e., a patient is represented as an embedding matrix which has a dimension of $N_c \times k$, where N_c is the number of medical concepts in the medical history of a patient and k is the dimension of all medical concept vectors. Usually, N_c varies from patient to patient. Given the temporal representations of pairwise patients, calculating the clinical similarity between pairwise patients is not intuitive. We will describe the patient similarity learning method in the following section.

2) SIAMESE CNN WITH SPP

The Siamese CNN model is a CNN-based architecture that usually contains two identical CNNs. To generate the fixed-length outputs from the temporal patient representations of arbitrary sizes, we replace the CNN with SPP-net. The twin SPP-nets have the same configuration with the same parameters and share weights. Two copies of this subnetwork are joined by a loss function at the top, which computes a patient similarity metric using the Euclidean distance between the patient vectors extracted by each subnetwork.

a: THE ARCHITECTURE OF SIAMESE CNN WITH SPP

Fig. 4 shows the architecture of patient similarity learning using Siamese CNN with SPP. Here, $\mathbf{X}_A = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K]^T$ and $\mathbf{X}_B = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N]^T$ are the representation matrices of two patients p_A, p_B respectively, where K and N are the lengths of two patient medical concept sequences, and $\mathbf{c}_i$ is the vector representation of medical concept c_i. We use $\mathbf{X}_A$ and $\mathbf{X}_B$ as the input to two identical SPP-nets with the same weights, and then the patient vectors $\mathbf{F_W}(\mathbf{X}_A)$ and $\mathbf{F_W}(\mathbf{X}_B)$ are extracted by each of the SPP-net that the Siamese network comprises. The output of Siamese CNN with SPP $Dis_{\mathbf{W}} = \|\mathbf{F_W}(\mathbf{X}_A) - \mathbf{F_W}(\mathbf{X}_B)\|$ measure the similarity between the patient vectors. Our hypothesis is that, on one hand, two patients of the same cohort will have the similar embedding vectors and therefore their distance is close to zero. On the other hand, two patients of different cohorts will have more different embedding vectors and therefore their distance will be larger.

The similarity between the feature vectors $\mathbf{F_W}(\mathbf{X}_A)$ and $\mathbf{F_W}(\mathbf{X}_B)$ of temporal patient matrices $\mathbf{X}_A$ and $\mathbf{X}_B$ can be measured by distance metrics such as those induced by the norms L_1 and L_2 or with similarity function such as cosine similarity. In our case, we choose Euclidean distance because it is widely used and have the best performance in a series of practices.

b: LOSS FUNCTION USED FOR SIAMESE CNN WITH SPP

Given $\mathbf{X}_A$ and $\mathbf{X}_B$ are a pair of input patient representations, $\mathbf{W}$ represents shared weighted matrix, and the mapping of $\mathbf{X}_A$ and $\mathbf{X}_B$ in the feature space is represented by $\mathbf{F_W}(\mathbf{X}_A)$ and $\mathbf{F_W}(\mathbf{X}_B)$, then Siamese CNN with SPP can be considered as a measure function that measures the similarity between $\mathbf{X}_A$ and $\mathbf{X}_B$, by calculating the Euclidean distance between the patient vectors. This learned similarity

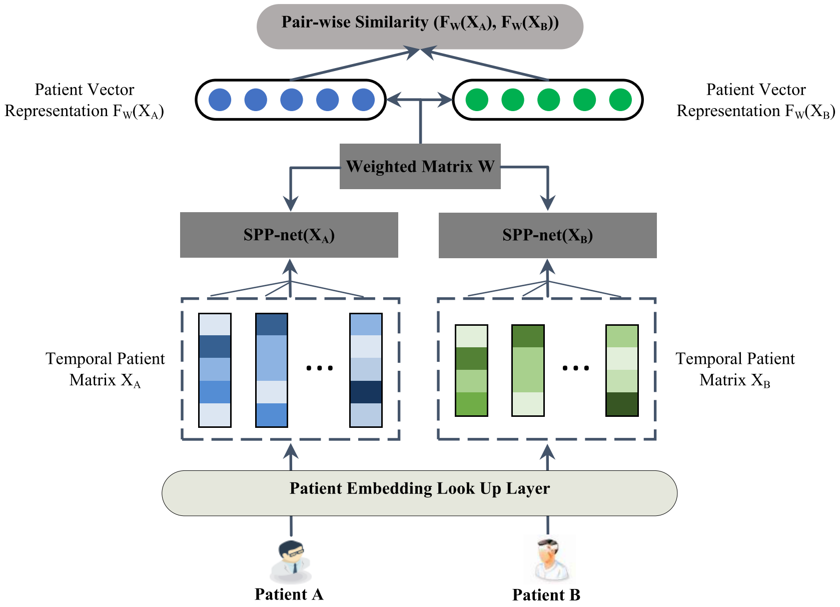

Pair-wise Similarity (F_W(X_A), F_W(X_B))
Patient Vector Representation F_W(X_A)
Patient Vector Representation F_W(X_B)
Weighted Matrix W
SPP-net(X_A)
SPP-net(X_B)
Temporal Patient Matrix X_A
Temporal Patient Matrix X_B
Patient Embedding Look Up Layer
Patient A
Patient B

FIGURE 4. Patient similarity learning using Siamese CNN with SPP.

measure function is defined as:

$$Dis_{\mathbf{W}}(\mathbf{X}_A, \mathbf{X}_B) = \|\mathbf{F_W}(\mathbf{X}_A) - \mathbf{F_W}(\mathbf{X}_B)\|_2 \tag{15}$$

During the training phase of Siamese CNN with SPP we use the contrastive loss function introduced by Chopra *et al.* in [45], which is defined as follows:

$$\begin{aligned} L(\mathbf{W}, Y, \mathbf{X}_A, \mathbf{X}_B) = & \frac{Y}{2} Dis_{\mathbf{W}}(\mathbf{X}_A, \mathbf{X}_B)^2 \\ & + \frac{1-Y}{2}(\max\{0, m - Dis_{\mathbf{W}}(\mathbf{X}_A, \mathbf{X}_B)\})^2 \end{aligned} \tag{16}$$

where $m > 0$ is a constant called a margin and Y is a binary label assigned to the pair of input patient representations, so that $Y = 1$ if the patients belong to the same cohort and $Y = 0$ otherwise.

Note that if the patients belong to the same cohort ($Y = 1$) their distance contributes to the loss function, while if they belong to different cohorts ($Y = 0$), only whose distance is less than or equal to m contribute. Therefore, minimizing $L(\mathbf{W}, Y, \mathbf{X}_A, \mathbf{X}_B)$ with respect to $\mathbf{W}$ would result in a small value of $Dis_{\mathbf{W}}(\mathbf{X}_A, \mathbf{X}_B)$ for patients of the same cohort and a large value of $Dis_{\mathbf{W}}(\mathbf{X}_A, \mathbf{X}_B)$ for patients of different cohorts.

D. ALGORITHM DESCRIPTION

Algorithm 1 represents our proposed framework for patient similarity learningŕDeepPS in detail. The inputs of DeepPS are a patient dataset, a set of patient's health records, a text corpus, a medical knowledge graph and the corresponding vocabulary of words, medical entities and relations. DeepPS has two main phrases. **Phrase 1:** After initializing medical entity/relation embeddings and projection matrix (Line 1), we alternatively learn from a batch of triplets randomly sampled from the medical knowledge graph (Line 3-6), and a batch of word pairs sampled by scanning the text corpus (Line 7-10). To combine such two information, we apply the alignment strategy based on the translation-based methods to join medical entity/relation and word embeddings into a unified low-dimensional semantic space (Line 11-18). **Phrase 2:** For each patient p, the temporal representation of patient p is denoted as $\mathbf{X}_p$ which is constructed by stacking all medical concepts vectors in the medical history of patient p (Line 20-21). Lastly, the patient matrices are fed into Siamese CNN with SPP to measure the similarity between all pair patients (Line 22). Additionally, for each patient, we select the patient corresponding to the highest similarity score (Line 23-29).

IV. EXPERIMENTS

In this section, the performance of the proposed framework, DeepPS, is evaluated using four real-world datasets. First, the experimental setup is introduced. Then we evaluate the effectiveness of proposed framework on three tasks: patient similarity analysis, patient clustering and visualization. In addition, we explore the influence of different metric functions in the patient similarity learning performance. Last, parameter sensitivity analyses of margin b, hyper-parameter γ, margin m and number of convolutional filters f are provided.

A. EXPERIMENTAL SETTINGS

1) DATASETS

Our experiments are performed on the real EHR dataset, MIMIC-III [46], and three knowledge bases ICD-9 ontology, DrugBank and Wikipedia which are publicly available in different forms.

- **MIMIC-III** (Medical Information Mart for Intensive Care III) is a large database of intensive care patients open to the public free of charge and collects all charted data (demographics, vital signs, medications,

Algorithm 1 Patient Similarity via Joint Embeddings of Medical Knowledge Graph and Medical Entity Descriptions

```
Input: A patient dataset $D$, a set of patient's health records $S = \{TC_1, TC_2, \dots\}$, a text corpus $\Delta$, a medical knowledge graph $G$, and the corresponding vocabulary of words, medical entities and relations ($V$, $E$ and $R$, respectively)
Output: A set of the most similar patients $\check{D}$
Initialize medical entity and relation embedding vectors $\mathbf{e}$ ($e\epsilon E$) and $\mathbf{r}$ ($r\epsilon R$), and projection matrix $\mathbf{H}_r$
repeat
  Sample a batch of triples from $G_{batch}$ from $G$
  for $(h, r, t)\epsilon G_{batch}$ do
    Update $\mathbf{h}$, $\mathbf{r}$, $\mathbf{t}$ by using Equation (1-4) with negative sampling
  end for
  Sample a batch of word pairs $\Delta_{batch}$ from $\Delta$
  for $(w, v)$ $\Delta$ $\Delta_{batch}$do
    Update $\mathbf{w}$ by using Equation (5-7) with negative sampling
  end for
  Sample a batch of medical entities $E_{batch}$ from $E$
  for $e\epsilon E_{batch}$ do
    Sample a batch of words $V_{batch}$ from entity description $text(e)$
    for $w\epsilon V_{batch}$ do
      Map $\mathbf{e}$ and $\mathbf{w}$ into a unified semantic space
      Update $\mathbf{e}$, $\mathbf{w}$ $\leftarrow$ based on Equation (8-11)
    end for
  end for
until Convergence
foreach $p$ $\epsilon D$ do
  Construct the patient matrix $\mathbf{X}_p \leftarrow$ based on $TC_p = \{c_1, c_2, \dots\}$
Train Siamese CNN with SPP using patient matrices as the input
$\check{D} \leftarrow \{\}$
foreach $p_i$ $\epsilon D$ do
  foreach $p_j$ $\epsilon D$ $\backslash p_i$ do
    Compute the similarity score between $p_i$ and $p_j$
  Rank the similarity score
  Select the patient $p_j$ corresponding to the highest similarity score
  $\check{D} \leftarrow p_j$
Return $\check{D}$
```

procedures, diagnoses, patient outputs, laboratory tests, physician notes, and treatment details) on ICU patients from Beth Israel Deaconess Medical Center between 2001 and 2012. The MIMIC-III dataset includes 6,918 distinct diseases, 4,525 distinct medicines and 2,003 distinct procedures from 46,297 unique patients.

- **ICD-9 ontology** (the 9th revision of the International Statistical Classification of Diseases and Related Health Problems) contains 13,000 international standard codes of diagnoses and their hierarchical relationships.
- **DrugBank** is a knowledge base containing extensive biochemical and pharmacological information about drugs, their mechanisms and their targets. It covers 7,683 active moieties. Although not primarily developed for clinical use, DrugBank provides a set of 12,128 drug-drug interactions (DDIs), asserted at the ingredient level, along with a brief textual description of the interaction, and information about the possible molecular basis of the interaction (target-based, enzyme-based, transporter-based).
- **Wikipedia** is a free multi-lingual online encyclopedia that is constructed in a collaborative effort of voluntary contributors and still grows exponentially. It contains more than 5.7 million articles and 46 million pages and is edited on average by more than 128k active users every month.

2) MEDICAL KNOWLEDGE GRAPH CONSTRUCTION

Before constructing a large medical knowledge graph by connecting MIMIC-III, ICD-9 ontology, and DrugBank, we introduce how to select patients and their medical concepts.

a: THE SELECTION OF PATIENTS

Following the cohort selection reasons in [47], we extract nine patient cohorts from the MIMIC-III dataset: Atherosclerosis, Heart Failure, Kidney Failure, Intestinal Diseases, Liver Diseases, Pneumonia, Septicemia, Respiratory Failure and Gastritis. To perform patient similarity learning, the patients from nine cohorts are selected as follows: (1) We remove the patients with missing data on admission date and discharge date; (2) We keep the patients which consist of at least three ICD-9 codes; (3) We remove the patients which have the discharge date after 2200/1/1; and (4) we remove the patients who suffer from more than one disease in the cohort list. Totally 26,009 patients are finally selected and divided into training set (80%), test set (10%) and validation set (10%).

b: THE SELECTION OF MEDICAL CONCEPTS

The medical knowledge graph consists of medical fact triples, where a medical fact triple indicates that a patient takes a medication, a patient performs a surgery, or a patient is diagnosed with a disease, etc. Therefore, we need to extract the diagnosis information, medication information, and procedure information from MIMIC-III as medical concepts of patients. A subset of medical concepts is selected by removing medical concepts which appear less than five patients to avoid biases and noise. Finally, there are totally 9,067 distinct medical concepts left.

After extracting patients and their medical concepts, we need to tackle the task of finding sameAs links between MIMIC-III and other biomedical knowledge graphs (ICD-9 ontology and DrugBank). In MIMIC-III, the ICD-9 codes for diagnoses and procedures can be directly linked to ICD-9 ontology by string matching. For the medication entity linking, medication names are various and often contain some insignificant words (10%, 200mg, glass bottle, etc.), which

TABLE 1. Statistics of medical entities.

Medical Entities	# Cardinality
Patient	26,009
Diagnosis	4,759
Medication	3,062
Procedure	1,246
Diagnosis-related	4,759
Procedure-related	1,278
Medication-related	1,500
Total	42,613

TABLE 2. Statistics of medical relations.

Relations	# Cardinality
Patient-Diagnosis	283,976
Patient-Medication	695,089
Patient-Procedure	95,698
Diagnosis-Diagnosis	6,037
Medication-Medication	36,768
Medication-Diagnosis	763,265
sameAs	8,117
Total	1,888,950

challenges the medication entity linking if the label matching method is directly used. In order to overcome this problem, we use an entity linking method which is mentioned in [48]. Table 1 and 2 show the statistics of the medical knowledge graph we construct. The medical knowledge graph will be used to learn low-dimension representations of medical entities and relations by the DeepPS framework.

3) MEDICAL ENTITY DESCRIPTION RETRIEVAL

The text corpus Δ is composed of medical entity descriptions from Wikipedia and EHRs. For patient entity descriptions, we use the discharge summary associated with each patient from MIMIC-III. For diagnosis and procedure entity descriptions, we use the detail description for the corresponding ICD-9 code in the tables from MIMIC-III. For remaining entity descriptions, we use the Wikipedia article associated with each corresponding entity. In contrast to the whole Wikipedia articles which might contain much noise, the summary section of each Wikipedia article generalizes the main topic and is of relatively higher quality. Therefore, the summary section of entity's corresponding Wikipedia article is also considered as its entity description in our work. Plus, the punctuation should be removed before being forwarded to training. Finally, we filter out rare words that appeared fewer than five times in the text corpus. Consequently, the total number of unique words in the text corpus is approximately 36,723.

4) COMPETING METHODS

To evaluate the effectiveness of the proposed DeepPS, we compare the framework with the following baselines and approaches in terms of different performance metrics.

- **Principal Component Analysis (PCA)**: A unsupervised method is widely used for dimension reduction and feature extraction [49]. We apply PCA on the one-hot EHR matrices of patients and perform Euclidean distance based on the PCA results.
- **Code Sum based Matching (CSM)**: A method presented by Choi *et al.* [50] that represents a patient by summing up all its ICD code vectors, absolutely eliminating the sequential structure of ICD codes. It determines the similarity between a pair of patients by computing the cosine distance between their summed vectors.
- **CNN_triplet**: A patient similarity learning framework proposed by Suo *et al.* [51] that uses CNN to capture local important information in EHRs and then feed the learned representation into triplet loss.
- **Deep Embedding**: A framework introduced by Chang *et al.* [47] that combines CNN with distributional medical event embeddings from Word2Vec. Based on the temporal embedding matrices, patient features are filtered through the convolutional layer of neural network. Feature maps that represent patient clinical characteristics are then used to measure the distance between patients.
- **TransR-DeepPS**: It applies TransR instead of the joint embedding model in the proposed DeepPS to learn medical entity/relation embeddings, without taking the additional corpus information into consideration.
- **T-DeepPS**: A Triplet architecture based on the proposed DeepPS inspired by Patient Similarity Deep Metric Learning Framework (PSDML).

5) EVALUATION METRIC

With generated representations of each patient, we calculate the similarity score among all patient pairs using two different criteria: hospital readmission rate and incident rate difference for mortality. With the inherent difficulty of measuring the patient similarity, these two criteria are chosen since (1) both hospital readmission rate and incident rate difference for mortality play a significant role in many patient matching applications [52] and (2) they are recorded in most routinely collected data, and hence have a broad prospect of application [53], [54]. Also, we use Rand Index [55] and Normalized Mutual Information [56] to evaluate the patient clustering. We will describe the detailed definitions of these four criteria next.

a: HOSPITAL READMISSION RATE (HRR)

Assume $Case = \{r_1, r_2, \ldots, r_N\}$ is the collection of readmission statuses of N patients and Control $= \{\check{r}_1, \check{r}_2, \ldots, \check{r}_N\}$ is the collection of readmission statuses of the most similar patients of N patients. HRR is computed as follows:

$$HRR = \sum_{i=1}^{N} \omega(Case[i], Control[i]) \tag{17}$$

where $\omega(Case[i], Control[i]) = \begin{cases} 0, Case[i] \neq Control[i] \\ 1, Case[i] = Control[i] \end{cases}.$

HRR measures the overall matching efficiency and $HRR \in [0, 1]$.

b: INCIDENCE RATE DIFFERENCE FOR MORTALITY (IRDM)

Assume $Case = \{(t_1, d_1), (t_2, d_2), \ldots, (t_N, d_N)\}$ is the collection of tuples (discharge date, death date) of N patients, where t_i is the discharge date, and d_i is the death date. The incidence rate of the collection of N patients is computed as follows:

$$IR(Case) = \frac{count(death)}{\sum\limits_{i=1,d_i \neq null}^{N} (d_i - t_i) + \sum\limits_{i=1,d_i = null}^{N} (d_{null} - t_i)} \tag{18}$$

where $count(death)$ is the number of patients which have the death dates and d_{null} is 2200/1/1.

Similarly, we can compute the incidence rate of the most similar patients of N patients, called $IR(Control)$. $IRDM$ is computed as follows:

$$IRDM = |IR(Case) - IR(Control)| \tag{19}$$

c: RAND INDEX (RI)

RI is the most frequently used evaluation metric in data clustering. *RI* is computed as follows:

$$RI = \frac{TP + TN}{\binom{n}{2}} \tag{20}$$

where *TP* is the number of times a pair of patients belonging to the same cohort who are grouped into one single cluster. *TN* is the number of times a pair of patients from different cohorts who are grouped into different clusters. n is the total number of patients. In general, the larger the value of *RI*, the more consistent the clustering results is with the real situation.

d: NORMALIZED MUTUAL INFORMATION (NMI)

NMI is often used in data clustering to measure the similarity of the two clustering results. *NMI* is computed as follows:

$$NMI(X, Y) = \frac{2 \cdot I(X, Y)}{[H(X) + H(Y)]} \tag{21}$$

where Mutual Information $I(X, Y)$ is the relative entropy of the joint distribution $p(x, y)$ and the product distribution $p(x)(y)$, whose formula is:

$$I(X, Y) = \sum_{x}\sum_{y} p(x, y) log \frac{p(x, y)}{p(x)p(y)} \tag{22}$$

$H(X)$ is the information entropy, and the formula is:

$$H(X) = -\sum_{i} p(x_i) log\ p(x_i) \tag{23}$$

Similar to the value of *RI*, the closer the value of *NMI* is to 1, the better is the patient clustering.

6) PARAMETER SETTINGS

a: JOINT EMBEDDING LEARNING

The configurations for joint embedding learning are given as follows: the learning rate α is set as 0.025, the hyper-parameter γ is among $\{0, 1E{-7}, 1E{-5}, 1E{-3}, 1E{-1}\}$, the margin b is among $\{2.0, 4.0, 6.0, 8.0, 10.0\}$, the dimensions of medical entity embedding k and relation embedding d are set as 100, the number of negative samples μ is set as 10, the context window size β is set as 5, the dimension of word embedding n is set as 100, and the distance function is set as L_1-norm.

TABLE 3. Hospital readmission rate (*HRR*).

Methods	Technique	*HRR*
PCA	Non-embedding	0.593
CSM	Word2Vec	0.684
CNN_triplet	CNN	0.722
Deep Embedding	Word2Vec-CNN	0.737
TransR-DeepPS	KGE	0.801
T-DeepPS	Triplet Loss	0.866
DeepPS	Jointly(desp)	0.879

TABLE 4. Incidence rate difference for mortality (1E-5).

Methods	Technique	*IRDM*
PCA	Non-embedding	0.420
CSM	Word2Vec	0.336
CNN_triplet	CNN	0.309
Deep Embedding	Word2Vec-CNN	0.298
TransR-DeepPS	KGE	0.261
T-DeepPS	Triplet Loss	0.241
DeepPS	Jointly(desp)	0.239

b: SIAMESE CNN WITH SPP

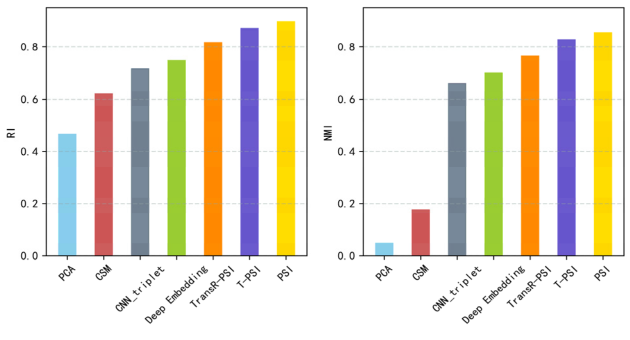
For Siamese CNN with SPP, we use the Adam [57] optimization algorithm as it is computationally efficient and exhibits faster convergence than standard stochastic gradient descent methods. The number of convolutional filters f is set to 50, 100, 150, 200, 250, and the margin m takes on 0.5, 1.0, 1.5, 2.0, 2.5. The metric function is among $\{L_1, L_2, \text{Cosine}\}$. The optimization hyper-parameters are fixed to $\gamma_1 = 0.9$, $\gamma_2 = 0.999$ with a learning rate of $\alpha = 0.0009$. In the SPP layer, we use a 3-level pyramid. The pyramid is $\{4 \times 4, 2 \times 2, 1 \times 1\}$ (totally 21 bins). We train Siamese CNN with SPP using 128 examples of shuffled mini-batches and adopt nonlinear rectification (ReLU) activation function. With regards to overfitting issue we add dropout regularization with dropout rate setting to 0.6.

B. EXPERIMENTAL RESULTS

1) PATIENT SIMILARITY ANALYSIS

We first compare the performance on the task of patient similarity analysis. We train the proposed DeepPS and all baseline methods to learn the similarity degrees among patients, and then use *HRR* and *IRDM* to measure the patient similarity. The results of *HRR* and *IRDM* are shown in Table 3 and 4 respectively.

As shown in Table 3 and 4, we observe that the values of *HRR* and *IRDM* of DeepPS are significantly higher than that of the baseline methods. More specifically, the proposed DeepPS achieves the best performance in *HRR* and *IRDM*, which are 0.879 and 0.239, respectively. Among all baseline methods, PCA, CSM, CNN_triplet and Deep Embedding achieve the lower values of *HRR* and *IRDM*. This is probably due to the fact that these methods learn lower dimensional feature representations directly from the correlation matrix or medical texts while not benefiting from the structural information brought by the medical knowledge graph. However, in the medical knowledge graph, the inner structure usually reflects the known medical facts, which could serve as important features for discriminating whether a pair of



(a) RI

(b) NMI

FIGURE 5. **Performance of patient clustering.**

patients are similar. TransR-DeepPS embeds entities and relations of a medical knowledge graph. It achieves fairly good results on patient similarity analysis although not as good as T-DeepPS. The superior performance of T-DeepPS indicates that medical knowledge graph embedding can be improved better through the external information such as medical entity descriptions. Compared with the best performing baseline T-DeepPS, the proposed DeepPS achieves an improvement from 0.866 to 0.879 in *HRR* and 0.241 to 0.239 in *IRDM*, which indicates that contrastive loss performs better than triplet loss on patient similarity analysis. Overall, DeepPS achieves the best results on large real datasets, demonstrating its generalizing ability in similarity learning of patients.

2) PATIENT CLUSTERING

Risk prediction can help the medical decision on identifying symptoms for early diagnosis, while patient clustering can help to analyze disease cohort distributions. The performance of patient clustering is shown in Fig. 5. We adopt k-means clustering algorithm with $k = 9$. In this experiment, we run the proposed DeepPS and all baseline methods to generate a representation vector for each patient, which is used as a feature representation for patient clustering.

We use *RI* and *NMI* to measure the performance of patient clustering. From Fig. 5, we can observe that the proposed DeepPS significantly outperforms the results of all baseline methods in patient clustering task. For instance, DeepPS increases *RI* from 20.2% compared with Deep Embedding to 44.3% compared with CSM; increases *RI* from 25.2% compared with CNN_triplet to 92.5% compared with PCA; increases *NMI* from 3.3% compared with T-DeepPS to 381.5% compared with CSM and increases *NMI* from 11.9% compared with TransR-PSI to 1638.3% compared with PCA. The wide margin in the results between DeepPS and all other baseline methods demonstrates the power of medical knowledge graph embedding and the importance of incorporating external information of medical entities. This experimental results on patient clustering can be used to further study of disease cohort distributions.

3) VISUALIZATION

We utilize the visualization tool *t-SNE* [58] to plot the low-dimensional patient representations learned by different patient similarity learning methods. As a result, each patient is mapped as a two-dimensional vector. Then we can visualize each vector as a point on a two-dimensional space. For patients which are labelled as different cohorts, we use different colors on the corresponding points. Therefore, a good visualization result is that the points of the same color are near from each other. The visualization figure is shown in Fig. 6.

From Fig. 6, we can see that the result of PCA is not satisfactory because the points belonging to different cohorts are mixed each other. For CSM, the clusters of different cohorts are formed. However, in the top part the patients of different cohorts are still mixed with each other and the boundaries of each group are not very clear. For CNN_triplet, the results look better because points of the same color form segmented groups. However, in the center part the patients of different cohorts are still mixed with each other. The result of Deep Embedding is better than that of CNN_triplet, which is because Deep Embedding uses the Skip-gram model to learn the embedding representations of medical concepts. Obviously, the visualization of DeepPS performs best in both the aspects of group separation and boundary aspects.

C. COMPARISONS BASED ON DIFFERENT METRIC FUNCTIONS

In this experiment, we first select two patient cohorts from the MIMIC-III dataset, namely, Pneumonia and Heart Failure. Then we additionally try to use other metric functions, such as the Manhattan distance (also known as L_1 distance) and the cosine similarity to analyze the two patient cohorts. Fig. 7 shows the performance of patient similarity learning by using different metric functions. As we can see, the L_2 distance as the Euclidean distance has the best effect on the two patient cohorts. Furthermore, the L_1 distance is better than the cosine similarity for the Heart Failure cohort and the cosine similarity is better than the L_1 distance for the Pneumonia cohort. The results indicate that using the L_1 (rather than L_2) distance can lead to undesirable plateaus in this case. That is mainly because we employ the contrastive loss function in this paper. Indeed, from the perspective of energy minimization, if the energy is the L_2 distance between the embedding vectors of the two patients, the gradient of the energy with respect to the parameter would vanish as the energy approached zero. This could lead to failure of the machine to learn in cases where the two patients are impostors and the corresponding energy is near zero. As such, depending on the differences of loss functions and patient cohorts, we should carefully choose the metric function.

D. PARAMETER SENSITIVITY ANALYSES

In the proposed framework DeepPS, there are two sets of parameters. One is the set of parameters for the joint embedding model, and the other set are the parameters for Siamese CNN with SPP. To analyze the influence of the parameters in the similarity learning performance of the proposed framework, we perform a parameter sensitivity evaluation for the four key parameters: the margin b, the hyper-parameter γ,

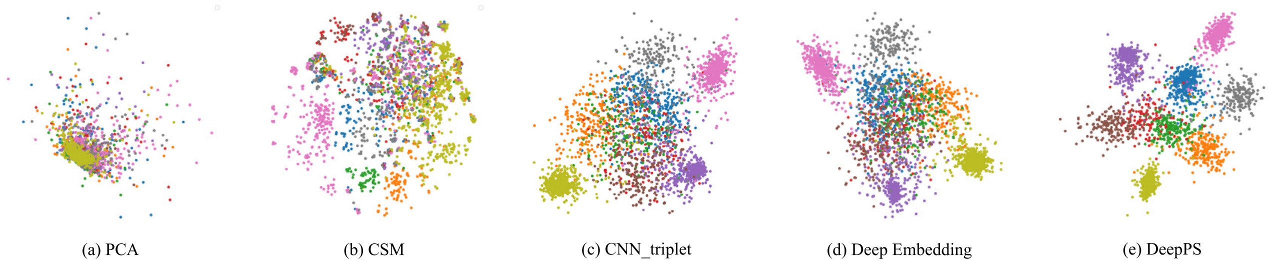

(a) PCA (b) CSM (c) CNN_triplet (d) Deep Embedding (e) DeepPS

FIGURE 6. Visualization of patients. Each point indicates one patient. Color of a point indicates the cohort of the patient.

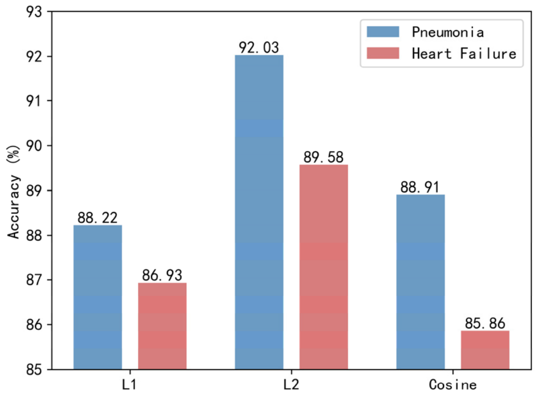
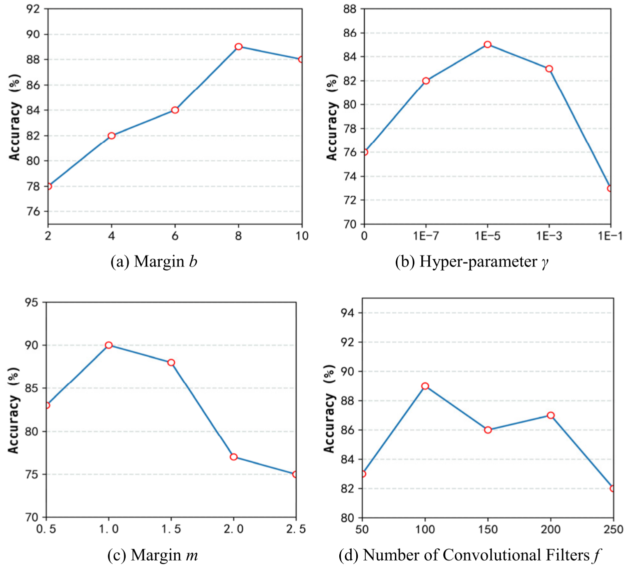


FIGURE 7. Performance of patient similarity learning using different metric functions.



(a) Margin b (b) Hyper-parameter γ

(c) Margin m (d) Number of Convolutional Filters f

FIGURE 8. Parameter sensitivity analyses.

the margin m and the number of convolutional filters f. The line plots in Fig. 8 show the accuracy of the proposed framework with different parameter values. In Fig. 8 (a), the margin b is used for the energy function of our joint embedding model, and the best performance is achieved when b is set as 8.0. From Fig. 8 (b), setting the hyper-parameter γ which is used for maximizing the objective function of our joint embedding model as 1E−5 shows the best performance for the proposed framework. The Fig. 8 (c)-(d) show the accuracy of the proposed framework in the margin m and the number of convolutional filters f. The margin m is used for the contrastive loss function of Siamese CNN with SPP and the number of convolutional filters f is used for the convolution operation of CNN. We find that $m = 1$ and $f = 100$ produce the best performance of the proposed framework.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel framework that learns the pair-wise patient similarity degree, referred to as DeepPS. To make full use of the semantics of medical entity descriptions, DeepPS enables a joint embedding model to learn simultaneously from medical knowledge triples that have been directly observed in a given medical knowledge graph, and medical entity descriptions which have rich semantic information about these medical entities. In such a way, our joint embedding model can perform better in knowledge graph representation learning, going beyond previous medical knowledge graph embedding methods. In addition, the learned embeddings can be used for patient similarity measuring leveraging Siamese CNN with SPP. Extensive experiments on real world datasets are conducted and demonstrate the effectiveness of DeepPS.

In future work, we will explore the following research directions: (1) We plan to extend the medical knowledge graph by considering more information of patients, such as vital signs, lab measurements and demographics. (2) There are various of information like textual information of relations or medical entity types. We will incorporate these information sources into our framework that jointly derives the latent representations for medical entities.

REFERENCES

[1] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk prediction with electronic health records: A deep learning approach," in *Proc. SIAM Int. Conf. Data Mining*, Jun. 2016, pp. 432–440, doi: 10.1137/1.9781611974348.49.

[2] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Towards personalized medicine: Leveraging patient similarity and drug similarity analytics," in *Proc. AMIA Joint Summits Transl. Sci.*, vol. 2014, 2014, pp. 132–136.

[3] X. Zhou, W. Liang, K. I.-K. Wang, and S. Shimizu, "Multi-modality behavioral influence analysis for personalized recommendations in health social media environment," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 5, pp. 888–897, Oct. 2019, doi: 10.1109/TCSS.2019.2918285.

[4] S.-H. Wang, Y. Zhang, Y.-J. Li, W.-J. Jia, F.-Y. Liu, M.-M. Yang, and Y.-D. Zhang, "Single slice based detection for Alzheimer's disease via wavelet entropy and multilayer perceptron trained by biogeography-based optimization," *Multimedia Tools Appl.*, vol. 77, no. 9, pp. 10393–10417, May 2018.

[5] S. Wang, J. Sun, I. Mehmood, C. Pan, Y. Chen, and Y. Zhang, "Cerebral micro-bleeding identification based on a nine-layer convolutional neural network with stochastic pooling," *Concurrency Comput., Pract. Exp.*, vol. 32, no. 1, p. e5130, Jan. 2020, doi: 10.1002/cpe.5130.

[6] Y.-D. Zhang, V. V. Govindaraj, C. Tang, W. Zhu, and J. Sun, "High performance multiple sclerosis classification by data augmentation and AlexNet transfer learning model," *J. Med. Imag. Health Informat.*, vol. 9, no. 9, pp. 2012–2021, Dec. 2019.

[7] Y. Zhang, S. Wang, Y. Sui, M. Yang, B. Liu, H. Cheng, J. Sun, W. Jia, P. Phillips, and J. M. Gorriz, "Multivariate approach for Alzheimer's disease detection using stationary wavelet entropy and predator-prey particle swarm optimization," *J. Alzheimer's Disease*, vol. 65, no. 3, pp. 855–869, Sep. 2018.

[8] S.-H. Wang, Y.-D. Zhang, M. Yang, B. Liu, J. Ramirez, and J. M. Gorriz, "Unilateral sensorineural hearing loss identification based on double-density dual-tree complex wavelet transform and multinomial logistic regression," *Integr. Comput.-Aided Eng.*, vol. 26, no. 4, pp. 411–426, Sep. 2019.

[9] S. Wang, C. Tang, J. Sun, and Y. Zhang, "Cerebral micro-bleeding detection based on densely connected neural network," *Frontiers Neurosci.*, vol. 13, p. 422, May 2019, doi: 10.3389/fnins.2019.00422.

[10] S.-H. Wang, S. Xie, X. Chen, D. S. Guttery, C. Tang, J. Sun, and Y.-D. Zhang, "Alcoholism identification based on an AlexNet transfer learning model," *Frontiers Psychiatry*, vol. 10, p. 205, Apr. 2019.

[11] X. Zhou, Y. Li, and W. Liang, "CNN-RNN based intelligent recommendation for online medical pre-diagnosis support," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, May 14, 2020, doi: 10.1109/TCBB.2020.2994780.

[12] W. Zhong, X. Yin, X. Zhang, S. Li, W. Dou, R. Wang, and L. Qi, "Multi-dimensional quality-driven service recommendation with privacy-preservation in mobile edge environment," *Comput. Commun.*, vol. 157, pp. 116–123, May 2020.

[13] X. Chi, C. Yan, H. Wang, W. Rafique, and L. Qi, "Amplified LSH-based recommender systems with privacy protection," *Concurrency Comput., Pract. Exper.*, Feb. 2020, doi: 10.1002/CPE.5681.

[14] H. Liu, H. Kou, C. Yan, and L. Qi, "Keywords-driven and popularity-aware paper recommendation based on undirected paper citation graph," *Complexity*, vol. 2020, pp. 1–15, Apr. 2020.

[15] L. Qi, Q. He, F. Chen, X. Zhang, W. Dou, and Q. Ni, "Data-driven Web APIs recommendation for building Web applications," *IEEE Trans. Big Data*, early access, Feb. 24, 2020, doi: 10.1109/TBDATA.2020.2975587.

[16] H. Liu, H. Kou, C. Yan, and L. Qi, "Link prediction in paper citation network to construct paper correlation graph," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–12, Dec. 2019.

[17] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. NIPS*, 2013, pp. 2787–2795.

[18] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proc. 28th AAAI*, Jun. 2014, pp. 1112–1119.

[19] Y. Lin, Z. Liu, X. Zhu, X. Zhu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. AAAI*, Feb. 2015, pp. 2181–2187.

[20] C. Kang, X. Yu, S.-H. Wang, D. Guttery, H. Pandey, Y. Tian, and Y. Zhang, "A heuristic neural network structure relying on fuzzy logic for images scoring," *IEEE Trans. Fuzzy Syst.*, early access, Jan. 13, 2020, doi: 10.1109/TFUZZ.2020.2966163.

[21] X. Zhou, W. Liang, K. I.-K. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep-Learning-Enhanced human activity recognition for Internet of healthcare things," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6429–6438, Jul. 2020, doi: 10.1109/JIOT.2020.2985082.

[22] L. M. Schriml, C. Arze, S. Nadendla, Y. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe, "Disease ontology: A backbone for disease semantic integration," *Nucleic Acids Res.*, vol. 40, no. 1, pp. 940–946, 2012.

[23] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, and A. Tang, "Drugbank 4.0: Shedding new light on drug metabolism," *Nucleic Acids Res.*, vol. 42, no. 1, pp. 1091–1097, 2014.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781

[25] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Sackinger, and R. Shah, "Signature verification using a siamese' time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, no. 4, pp. 669–688, 1993.

[26] S.-H. Wang, J. Sun, P. Phillips, G. Zhao, and Y.-D. Zhang, "Polarimetric synthetic aperture radar image segmentation by convolutional neural network using graphical processing units," *J. Real-Time Image Process.*, vol. 15, no. 3, pp. 631–642, Oct. 2018.

[27] J. Lee, D. M. Maslove, and J. A. Dubin, "Personalized mortality prediction driven by electronic medical data and a patient similarity metric," *PLoS ONE*, vol. 10, no. 5, May 2015, Art. no. e0127428.

[28] J. Sun, D. Sow, J. Hu, and S. Ebadollahi, "Localized supervised metric learning on temporal physiological data," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 4149–4152, doi: 10.1109/ICPR.2010.1009.

[29] D. Nguyen, W. Luo, S. Venkatesh, and D. Phung, "Effective identification of similar patients through sequential matching over ICD code embedding," *J. Med. Syst.*, vol. 42, no. 5, p. 94, May 2018.

[30] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, "Towards heterogeneous temporal clinical event pattern discovery: A convolutional approach," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 453–461, doi: 10.1145/2339530.2339605.

[31] Z. Che, Y. Cheng, Z. Sun, and Y. Liu, "Exploiting convolutional neural network for risk prediction with medical feature embedding," 2017, *arXiv:1701.07474*. [Online]. Available: http://arxiv.org/abs/1701.07474

[32] F. Zhao, J. Xu, and Y. Lin, "Similarity measure for patients via a siamese CNN network," in *Proc. Int. Conf. Algorithms Archit. Parallel Process.*, 2018, pp. 319–328, doi: 10.1007/978-3-030-05054-2_25.

[33] A. Callahan, J. Cruz-Toledo, P. Ansell, and M. Dumontier, "Bio2rdf release 2: Improved coverage, interoperability and provenance of life science linked data," in *The Semantic Web: Semantics Big Data*. Berlin, Germany: Springer, 2013, pp. 200–212, doi: 10.1007/978-3-642-38288-8_14.

[34] B. Chen, X. Dong, D. Jiao, H. Wang, Q. Zhu, Y. Ding, and D. J. Wild, "Chem2Bio2RDF: A semantic framework for linking and data mining chemogenomic and systems chemical biology data," *BMC Bioinf.*, vol. 11, no. 1, p. 255, 2010.

[35] D. J. Odgers and M. Dumontier, "Mining electronic health records using linked data," in *Proc. AMIA Joint Summits Transl. Sci.*, 2015, pp. 217–221, 2015.

[36] J. Pathak, R. C. Kiefer, and C. G. Chute, "Applying linked data principles to represent patient's electronic health records at mayo clinic: A case report," in *Proc. 2nd ACM SIGHIT Symp. Int. health Informat. (IHI)*, 2012, pp. 455–464, doi: 10.1145/2110363.2110415.

[37] V. N. Slee, "The international classification of diseases: Ninth revision (ICD-9)," *Ann. Internal Med.*, vol. 88, no. 3, pp. 424–426, 1978.

[38] N. Listed, "National drug code directory," *Hospitals*, vol. 43, no. 19, p. 80, 1969.

[39] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph and text jointly embedding," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Doha, Qatar: Association Computing Linguistics, 2014, pp. 1591–1601, doi: 10.3115/v1/D14-1167.

[40] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, 2010, pp. 177–186, doi: 10.1007/978-3-7908-2604-3_16.

[41] C.-H. Shih, B.-C. Yan, S.-H. Liu, and B. Chen, "Investigating siamese LSTM networks for text categorization," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 641–646, doi: 10.1109/APSIPA.2017.8282104.

[42] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833, doi: 10.1007/978-3-319-10590-1_53.

[43] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655, doi: 10.1097/00003643-201406001-00333.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[45] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.

[46] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, Dec. 2016, Art. no. 160035.

[47] Z. Zhu, C. Yin, B. Qian, Y. Cheng, J. Wei, and F. Wang, "Measuring patient similarities via a deep architecture with medical concept embedding," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 749–758, doi: 10.1109/ICDM.2016.0086.

[48] M. Wang, J. Zhang, J. Liu, W. Hu, S. Wang, X. Li, and W. Liu, "PDD graph: Bridging electronic medical records and biomedical knowledge graphs via entity linking," 2017, *arXiv:1707.05340*. [Online]. Available: http://arxiv.org/abs/1707.05340

[49] I. T. Jolliffe, "Principal component analysis," *J. Marketing Res.*, vol. 87, p. 513, Aug. 2002.

[50] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Medical concept representation learning from electronic health records and its application on heart failure prediction," 2016, *arXiv:1602.03686*. [Online]. Available: http://arxiv.org/abs/1602.03686

[51] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, J. Gao, and A. Zhang, "Deep patient similarity learning for personalized healthcare," *IEEE Trans. Nanobiosci.*, vol. 17, no. 3, pp. 219–227, Jul. 2018.

[52] J. P. Bigus *et al.*, "Information technology for healthcare transformation," *IBM J. Res. Develop.*, vol. 55, no. 5, pp. 492–505, Sep./Oct. 2011.

[53] Y. Wang, M. M. Pandolfi, J. Fine, M. L. Metersky, C. Wang, S.-Y. Ho, D. Galusha, S. V. Nuti, K. Murugiah, A. Spenard, T. Elwell, and H. M. Krumholz, "Community-level association between home health and nursing home performance on quality and hospital 30-day readmissions for medicare patients," *Home Health Care Manage. Pract.*, vol. 28, no. 4, pp. 201–208, Nov. 2016.

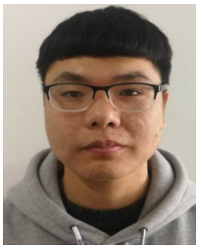
[54] S. J. Håkonsen, P. U. Pedersen, M. Bjerrum, A. Bygholm, and M. D. J. Peters, "Nursing minimum data sets for documenting nutritional care for adults in primary healthcare: A scoping review," *Int. J. Evidence-Based Healthcare*, vol. 16, no. 1, pp. 117–139, Jan. 2018.

[55] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, Dec. 1971.

[56] M. Meil, "Comparing clusterings—An information based distance," *J. Multivariate Anal.*, vol. 98, no. 5, pp. 873–895, 2007.

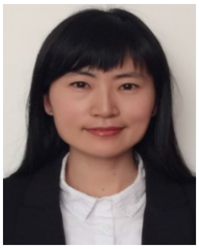
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[58] S. Arora, W. Hu, and P. Kothari, "An analysis of the t-SNE algorithm for data visualization," in *Proc. Conf. Learn. Theory*, 2018, pp. 1455–1462.



ZHIHUANG LIN is currently pursuing the master's degree with the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, China. He was with the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, from 2014 to 2018. His current research interests include deep learning, network embedding, and data mining.

DAN YANG received the M.S. and Ph.D. degrees in computer software and theory from Northeastern University, China, in 2004 and 2013, respectively. She was a Visiting Scholar with the New Jersey Institute of Technology, USA, from June 2015 to May 2016, supported by the Chinese Scholarship Council of the Ministry of Education. She is currently a Professor with the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, China. Her research interests include data integration, big data management, and applications in health care. She is a member of the China Computer Federation (CCF).



XIAOCHUN YIN received the B.S. degree in education and technology from Qufu Normal University, Qufu, China, in 2004, the M.S. degree in education and technology from Nanjing Normal University, Nanjing, China, in 2007, and the Ph.D. degree from Dongseo University, South Korea, in 2015. She is currently working as an Associate Professor with Weifang University of Science and Technology, China. Her research interests include network security, the IoT security, authentication protocol, and agricultural intelligence systems.