Patient outcome prediction via convolutional neural networks based on multi-granularity medical concept embedding

Conference Paper · November 2017

DOI: 10.1109/BIBM.2017.8217753

Yujuan Feng
3 PUBLICATIONS **35** CITATIONS
SEE PROFILE

Xiaolei Xie
Tsinghua University
51 PUBLICATIONS **522** CITATIONS
SEE PROFILE

Haibo Wang
Peking University
127 PUBLICATIONS **1,489** CITATIONS
SEE PROFILE

Ting Chen
Tsinghua University
244 PUBLICATIONS **8,525** CITATIONS
SEE PROFILE

Patient Outcome Prediction via Convolutional Neural Networks based on Multi-Granularity Medical Concept Embedding

Yujuan Feng [*†], Xu Min [*†], Ning Chen [*†], Hu Chen [††], Xiaolei Xie [‡¶], Haibo Wang [§‖¶] and Ting Chen [*†¶]
[*] MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST;
[†] Department of Computer Science and Technology; State Key Lab of Intelligent Technology and Systems;
[‡] Department of Industrial Engineering, Tsinghua University, Beijing 100084, China;
[§] Clinical Trial Unit, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, Guangdong 510080, China;
[††] Bureau of Medical Administration National Health and Family Planning Commission People's Republic of China;
[‖] China Standard Medical Information Research Center, Shenzhen, Guangdong 518054, China;
[¶] Corresponding authors. Email: xxie@tsinghua.edu.cn, haibo.wang@hqms.org.cn, tingchen@tsinghua.edu.cn

***Abstract*—The large availability of biomedical data brings opportunities and challenges to health care. Representation of medical concepts has been well studied in many applications, such as medical informatics, cohort selection, risk prediction, and health care quality measurement. In this paper, we propose an efficient multichannel convolutional neural network (CNN) model based on multi-granularity embeddings of medical concepts named MG-CNN, to examine the effect of individual patient characteristics including demographic factors and medical comorbidities on total hospital costs and length of stay (LOS) by using the Hospital Quality Monitoring System (HQMS) data. The proposed embedding method leverages prior medical hierarchical ontology and improves the quality of embedding for rare medical concepts. The embedded vectors are further visualized by the t-Distributed Stochastic Neighbor Embedding (t-SNE) technique to demonstrate the effectiveness of grouping related medical concepts. Experimental results demonstrate that our MG-CNN model outperforms traditional regression methods based on the one-hot representation of medical concepts, especially in the outcome prediction tasks for patients with low-frequency medical events. In summary, MG-CNN model is capable of mining potential knowledge from the clinical data and will be broadly applicable in medical research and inform clinical decisions.**

***Keywords*-electronic health records; multi-granularity embedding; convolutional neural network; hierarchical ontology;**

I. Introduction

Recently, China has been strengthening its health care system through reform and open policy of health care, focusing on hospital capacity [1] and the insurance coverage [2]. Despite having the largest population with over 1.3 billion, China has disproportionally fewer large-scale clinical studies and under-produces medical knowledge compared with other countries. Tremendous available multi-source biomedical data including electronic health records (EHR), imaging and -omics are produced in an explosive rate from claims and cost data, pharmaceutical and R&D data, and clinical data. In addition, medical ontologies used for data generalization (e.g. International Classification of Disease-10th version (ICD-10) [3], Unified Medical Language System(UMLS) [4], Medicine-Clinical Terms (SNOMED-CT) [5]) and existing expert knowledge (e.g. online medical encyclopedia and PubMed abstracts) should be mined to extract reliable content for healthcare research. It is necessary and urgent to leverage existing clinical resources to address the gaps in quality of care and mine high impact knowledge to inform clinical decisions and national policy [6].

Feature engineering are crucial and challenging in various fields such as image processing [7], [8] , language modeling [9], [10], as well as healthcare analytics [11], [12] due to complex, high dimensional and heterogeneous biomedical data. Traditional handcraft feature extraction and one-hot representation indicating features with categorical variables [13] scale poorly and are ineffective to uncover novel patterns from medical data. Deep learning methods are good representation-learning algorithms and have been well applied in healthcare domain recent years. [14]. For example, Cheng *et al*. [15] applied convolutional neural network (CNN) for prediction of congestive heart failure and chronic obstructive pulmonary disease. Nguyen *et al*. [16] proposed deeper, an end-to-end deep learning system based on CNNs to predict unplanned readmission after discharge. Che *et al*. [17] used stacked denoising autoencoders (SDAs) regularized with a prior knowledge based on ICD-9 to detect characteristic patterns of physiology.

Deep neural language models were used to learn embedded representation of medical concepts, such as diagnoses, medications, surgeries, and laboratory tests, which are generated from different materials. For example, Minarro-Gimenez *et al*. [18] developed a method to learn embedding from unstructured medical corpus crawled from PubMed, Merck Manuals, Medscape and Wikipedia. De Vine *et al*. [19] first extracted UMLS concepts from two sets of free texts, clinical patient records and medical journal abstracts and then learned the embedding using documents obtained by concatenating all of the extracted concepts. Choi *et al*. [13] learned medical concept representation from EHR for the heart failure prediction. Choi *et al*. [20] also proposed a methodological framework

978-1-5090-3050-7/17/$31.00 ©2017 IEEE

to learn low-dimensional representations for a wide range of medical concepts from different texts, including medical journals, medical claims and clinical narratives. Tran *et al*. [12] used RBMs to learn abstractions of ICD-10 codes to predict suicide risk for mental health patients.

In this paper, we develop a deep learning framework named MG-CNN, to estimate the association between patients characteristics and resource use including total hospital costs and LOS. It is a convolutional neural network based on multi-granularity embedding of medical concepts from EHR data. Our multi-granularity embedding method utilizes the prior medical ontology information based on ICD system and encodes medical concepts into a low-dimensional space, by using the Skip-gram algorithm. It exploits the co-occurrence information to conserve the relatedness of concepts. We show that the multi-granularity embedding method improves the quality of embedding for rare medical concepts and overcome the limitation of high dimensionality and sparse problem. The embedded vectors of all medical concepts are visualized in a 2-D space by the t-Distributed Stochastic Neighbor Embedding (t-SNE) technique [21], showing the expected property of grouping related concepts well. Besides, we compare the performance with traditional one-hot based regression methods such as linear regression and random forest, the MG-CNN model has superior effectiveness in patient outcome prediction tasks especially for the patients with rare medical events.

II. Methods

A. Preliminary

Learning distributed representation has been proven useful in natural language processing. The word2vec model [22], [23] is a neural network-based model for word embedding, which means mapping the word into a dense, real-valued vector space. The key idea is to embed semantically similar words, such as ‘enjoy’ and ‘like’, into adjacent real-value vectors in a Euclidean space, since they often occur in similar contexts.

Let $C = \{c_1, c_2, \dots, c_N\}$ be a collection of N medical concepts. Our goal is to learn a set of D-dimensional vector representations $V = \{v_1, v_2, \dots, v_N\}, v_i \in \mathbb{R}^D$ corresponding to C. In other words, we aim to embed N medical concepts into a Euclidean space $\mathbb{R}^D$, where N is usually very large, while D is usually set as 50-1000.

The key principle of the Skip-gram model in word2vec as illustrated in Fig. 1, is to learn vector representations which are capable of predicting nearby words given the center word within one sentence. Formally, given a sentence, i.e., a sequence of concepts $c_{i_1}, c_{i_2}, \dots, c_{i_T}$, $i_t \in [1, N]$, $t = 1, \dots, T$, T is the total number of concepts in ith sentence. The Skip-gram model maximizes average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-w\leq j-t\leq w, j\neq t} p(c_{i_j} \mid c_{i_t}), \tag{1}$$

where w is the window size of the training context, and the basic conditional probability $p(c_{i_j} \mid c_{i_t})$ is defined using softmax function:

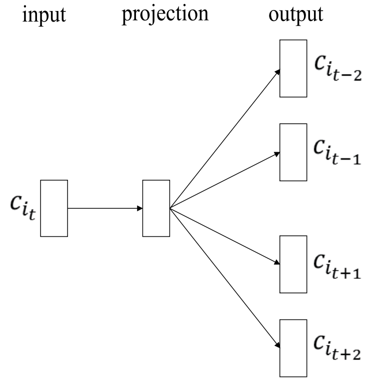

Fig. 1: The architecture of Skip-gram model.

$$p(c_k \mid c_s) = \frac{exp(v_k^T v_s)}{\sum_{v\in V} exp(v^T v_s)}, \tag{2}$$

where v_k and v_s are the embedding vectors of concepts c_k and c_i, respectively, and $k = i_j, s = i_t$, $k \in [1, N], s \in [1, N]$. In neural probabilistic language models, embeddings are viewed as parameters and trained by minimizing the loss function.

B. Construction of Medical Sentences

We put forward four strategies to generate sentences as the input of embedding method according to data characteristics. Within one inpatient record, we have one main diagnosis (ICD-10 code) and, at most ten secondary diagnoses (ICD-10 codes) with, at most ten procedure codes (ICD-9 codes) as depicted in Fig. 2. The diagnoses and procedures are denoted by the International Classification of Disease-Version 10 (RC020 ICD-10, Beijing Version) diagnostic codes [24] and International Classification of Disease-Version 9 (RC022 ICD-9, Beijing Version) procedure codes [25], respectively. Moreover, the standardized ICD codes are widely used as a billing and reimbursement system at hospitals.

The Strategy 1 (naive) simply concatenates diagnosis codes and procedure codes sequentially in the original order. However, if we assume that the main diagnosis and the first three secondary diagnoses are more reliable than the remaining secondary diagnoses, then Strategy 2, or permutation, can set the main diagnosis at the center of the sentence surrounded by the first three secondary diagnoses, followed by arranging the remaining secondary diagnoses randomly toward the two ends. Strategy 3, or shuffle, continues the idea in Strategy 2, except that we randomly shuffle all concepts instead of permuting them in a specific order. Finally, Strategy 4, or pairs, samples code pairs randomly to construct one sentence. Since the Skip-gram model only depends on pairs of nearby words [22], we duplicate pairs among the main diagnosis and the first three secondary diagnoses in order to emphasize the important pairs. We summarize four construction strategies in Fig. 2.

C. Multi-granularity Medical Concept Embedding

The standardized ICD codes are organized with certain rules and its hierarchical ontology contains abundant prior

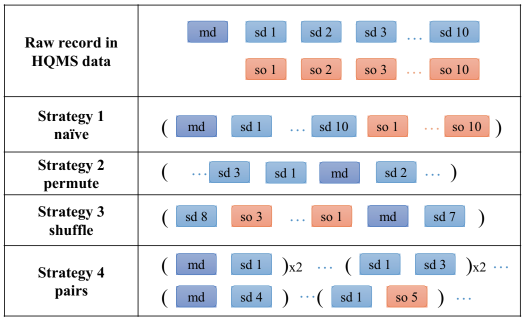
Raw record in HQMS data
md
sd 1
sd 2
sd 3
…
sd 10
so 1
so 2
so 3
…
so 10
Strategy 1 naïve
(md sd 1 … sd 10 so 1 … so 10)
Strategy 2 permute
(… sd 3 sd 1 md sd 2 …)
Strategy 3 shuffle
(sd 8 so 3 … so 1 md sd 7)
Strategy 4 pairs
(md sd 1)x2 … (sd 1 sd 3)x2 …
(md sd 4) …(sd 1 so 5) …

Fig. 2: Four sentence construction strategies based on raw record. The rectangular blocks contain the terms md, sd, and so, representing main diagnosis, secondary diagnosis, and procedure, respectively.

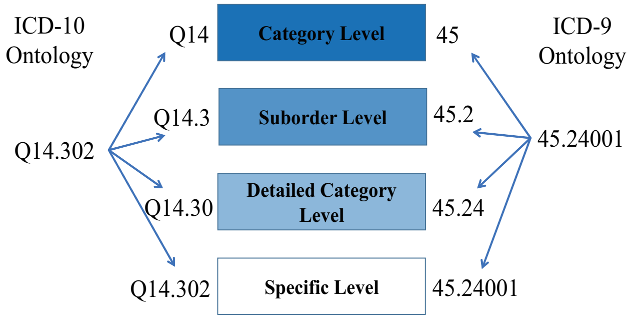

Fig. 3: Overview of multi-granularity embedding method. There are four taxonomy levels in the ICD-9 and ICD-10 classification system, and the ontology is hierarchical.

information. ICDs structure starts from the basic and leads to the specific, with each character of code implying special insight of disease or procedure characteristics [26]. There are four taxonomy levels of codes in both the ICD-9 code classification system and ICD-10 code classification system, and codes in each taxonomy level contain particular number of characters.

In the HQMS data, we can extract four fine-grained codes from each original ICD code. It is a way of data augmentation and the multi-granularity extraction of ICD code is shown in Fig. 3. For code with long characters in the specific form, such as Q14.302 whose description is congenital malformation of posterior segment of eye, now we can increase its occurrence frequency by counting the shorter codes constructed by first several characters in high taxonomy level. That are ‘Q14’, ‘Q14.3’, and ‘Q14.30’, ‘Q14.302’ corresponding to three-character categories (Category level), four-character subcategories (Suborder level), five-character subcategories (Detailed category level) and specific ICD-10 diagnosis code (Specific level), respectively. Similarly, for the ICD-9 procedure code 45.24001 whose description is inspection of lower intestinal tract via natural or artificial opening endoscopy, there are four fine-grained codes, i.e. ‘45’, ‘45.2’, ‘45.24’, ‘45.24001’, corresponding to two-character categories (Category level), three-character subcategories (Suborder level), four-character subcategories (Detailed category level) and specific ICD-9 procedure code (Specific level), respectively. This is of great benefits to understanding the rare medical concepts, which occur with extremely low frequency in HQMS data. The multi-granularity embedding of the medical concepts incorporates the prior information encoded in the ontology of the ICD structure. After the ICD code augmentation, we accordingly obtain four types of sentences composed of ICD codes in four levels. Then these sentences are fed to Skip-gram model to learn embeddings for codes in each level. At last, we can obtain four levels of embedding vectors corresponding to a specific ICD code in the raw HQMS data.

D. Our CNN Architecture

The proposed model architecture is depicted in Fig. 4. In the MG-CNN model, the input layer contains three types of inputs of shape $1{\times}L$, where L is length of the input ICD-9/ICD-10 codes sequences or demographic features. In the multi-granularity embedding layer, the input ICD codes sequences are initialized with 4-channel embedding matrixes of shape $4 \times L \times D$, where D is the embedding length of ICD code and set to be 600 in our experiment. The convolutional layer contains multiple kernels of variable size to extract relations between the medical concepts in different perspective, i.e. 200 kernels for each shape of $3 \times D$, $4 \times D$, $5 \times D$. Each kernel is applied to all the channels and the results are added to calculate intermediate medical code representation, the model is otherwise equivalent to the single channel architecture. Then global max-pooling is applied to select the most discriminative feature for each feature map and deals with variable lengths of inputs. We concatenate the demographic information vector with the intermediate medical code vectors to obtain the final patients representation. These features form the penultimate layer and are passed to three fully connected layers of size 1000, 500 and 100 with the rectified linear unit (ReLU) activation function between them. The output is the predicted value of target variable. The loss function to be minimized for our regression network is mean squared error (MSE). The dropout ratio is set 0.2 to avoid overfitting.

E. Patient Representation

We implement several ways to construct patients representation. We obtain a patients medical vector (D-dim) by simply aggregating all the medical concept vectors of this patient. Besides, we can learn a dense, low-dimensional patient representation through the CNN model. In addition, we extend the patient representation by concatenating the demographic vector (M-dim) and medical vector (D-dim), resulting to a final representation of a $(M + D)$-dim vector, where M is the number of elements in the demographic information.

F. Prediction of Outcome

It is the underlying assumption that, in general, patients with more complex complications would stay in the hospital longer and use more services and therefore generate greater hospital costs [27]. We compare the MG-CNN model and traditional models based on one-hot representation in the prediction of

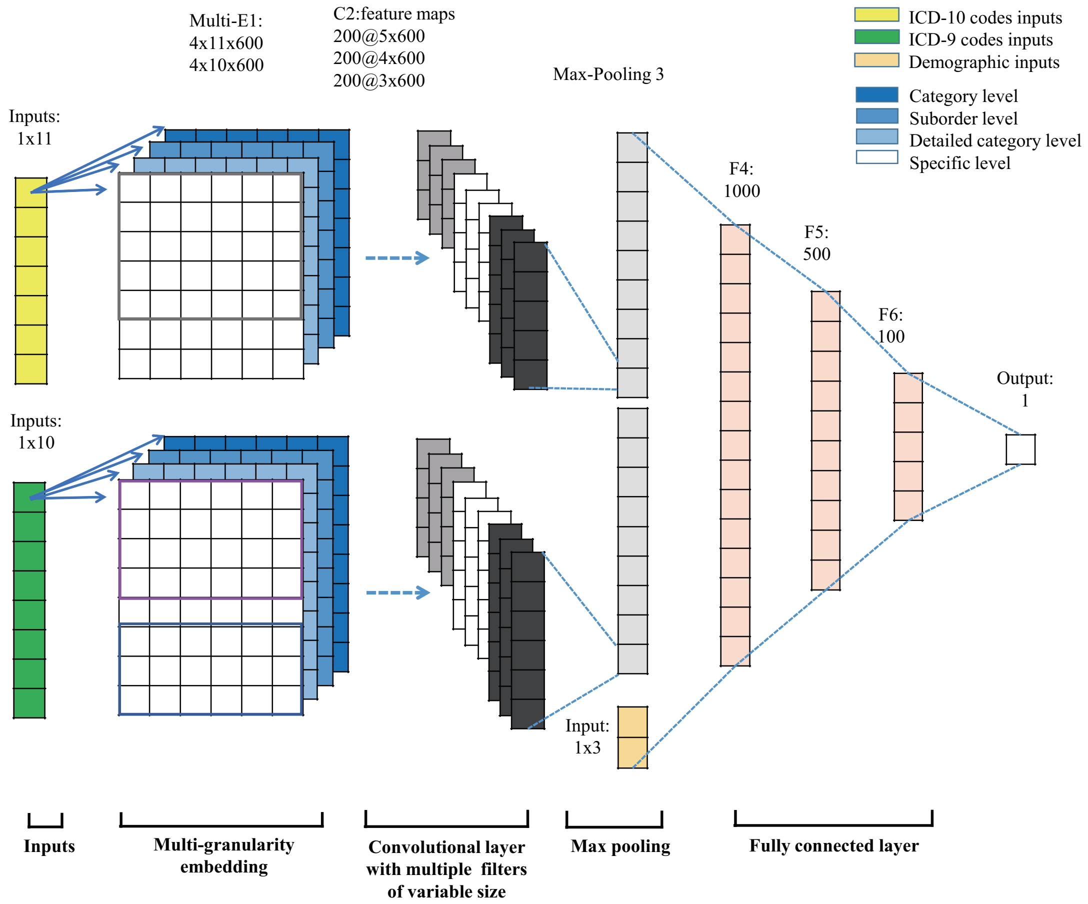

Fig. 4: Overview of proposed MG-CNN model. The ICD-10 codes, ICD-9 codes and demographic inputs refer to diagnosis codes, procedure codes and demographic information, respectively. In the multi-granularity embedding layer, diagnostic code and procedure code are initialized with embedding vectors in four taxonomy levels. Three types of convolutional kernel of different size (C2: feature maps) are used to recognize variant patterns of inputs. The demographic inputs are concatenated with the high-level representation of medical concepts. The parameter details of architecture are noted on the top of layers.

patients outcome. The outcome variables studied include the total hospital costs and LOS, as indicators of disease severity and resource utilization. We assess performance through prediction accuracy, as measured by the R-squared statistic in

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{3}$$

where $\{y_i\}$ and $\{f_i\}$ are the observed and predicted data, respectively, and $\bar{y}$ is the mean of the observed data. As R^2 approaches to 100%, the model incrementally improves its ability to predict the target variables.

III. Experiments

A. Data Source and Data Preprocessing

The Hospital Quality Monitoring System (HQMS) [28], the first official clinical data collection system of China launched in 2011, is designed to electronically collect face sheet information in all inpatient medical records from tertiary hospitals every day. We use hospital records of nearly $800,000$ inpatients dated from Jan 2013 to Dec 2015, and one patient record contains medical events (clinical diagnoses, surgical operations), demographic characteristics (age, gender, number of admissions) and outcome indicators like length of stay, total hospital costs etc.

Preprocessing is necessary for accurate predictive model. Firstly, we simply filter out records without complete demographic characteristics, length of stay and total hospital costs. Secondly, to avoid experimental error, we remove outliers whose total hospital costs are out of range from the $3rd$ to the $97th$ percentile of all samples' total hospital costs. Likewise, outliers whose LOS are out of the bound from $3rd$ to the $97th$ percentile of total samples' LOS are excluded. $550,704$ patients are available for subsequent analysis after exclusion. The average of LOS is 8.2 ± 5.6 days (range, $[1, 28]$ days), and the mean of total hospital costs is 18213.47 ± 18777.56 yuan (range, $[1406.35, 94735.72]$ yuan), showing large variances of target variables. The deep CNN model only accepts ICD code sequences with fixed length while patients vary in number of ICD codes. To tackle this issue, we set the maximum number of codes as the the input length for ICD-10 and ICD-9

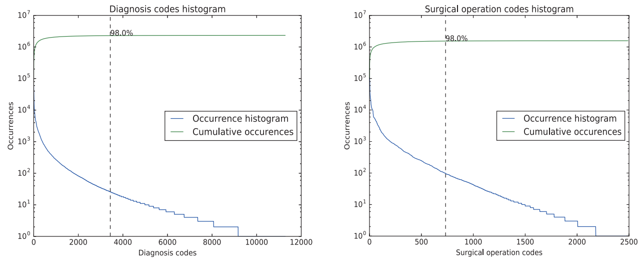

Fig. 5: Summary of diagnostic codes and procedure codes in HQMS. The threshold is set as 98.0% in the cumulative occurrences to divide codes into rare codes and common codes.

codes sequences, where $L(\text{ICD10}) = 11$ and $L(\text{ICD9}) = 10$, respectively. For patients with less than L codes, we impute with zero in the end of sequence. And the demographic length is $M = 3$.

B. Multi-granularity Medical Concept Embedding

There are total 15216 unique ICD codes of specific level, including diagnostic codes and procedure codes. With the multi-granularity medical concept extraction, we finally get 1474 unique ICD codes of the Category level, 5497 of the Suborder level, and 7619 of the detailed category level. The multi-granularity medical concept embedding increases the quality of embedding for rare code by increasing its frequency of occurrence. As for each ICD code in the record, we finally learn four channels of embedding vectors through skip-gram with Strategy 3 to construct sentences. Four different sentence construction strategies noted above are tested and compared in the experiment, Strategy 3 that shuffles all the diagnostic codes and procedures codes in one sentence yields the best performance and Strategy 1 has the worst performance. Strategy 2 and Strategy 4 produce similar prediction accuracy.

To prove the advantage of the multi-granularity embedding method, we select two special datasets constructed by 'rare patients' who have more than two even four rare ICD codes. Then we compare the performance of different predictive models on these rare patient datasets with low frequency medical events. Based on the observation in Fig .5, we find that a small subset of codes account for a large fraction of the total occurrences of data. The codes are ordered in terms of occurrences in cumulative analysis. Then we intuitively select codes retaining 98% occurrences of the raw data as the common codes, and the rest as rare codes. This results in 3435 unique common diagnostic codes and 7845 rare diagnosis code. Similarly, we get 1250 unique common procedure codes and 3236 rare procedure codes. In detail, we firstly select out 521 patients who have at least 4 rare codes from all the $550,704$ patients, we call it as 'extremely rare dataset'. Secondly, we randomly select 80% of samples for training, and the rest of 20% samples for testing, we call it as 'general dataset'. Thirdly, we select patients with at least 2 rare codes from 'general dataset' and is named as the 'rare dataset'. The experiments are carried out for several times for valid results.

C. Model Summary

Linear Regression with the sum of one-hot vectors (one-hot+): This is the baseline model. We filter all the rare codes to avoid the memory error caused by too sparse and high dimensional representation. The filtering reduces the feature space from 15216-dimensional to 4685-dimensional. Then the patient's representation is got by simply summing all the one-hot vectors of patients ICD codes and then concatenating with the demographic features.

Random Forest with the sum of one-hot vectors (one-hot+): This is also the baseline model. It is to be noted that only two baseline models have the filtering of rare codes. Because the embedding method and CNN model are capable of mapping the spares, high dimensional vector space to the low dimensional vector space.

Random Forest with sum of single channel word2vec embeddings (single channel Word2vec+): The patient's representation is obtained by simply summing all the embedding vectors of ICD codes in the Specific level.

Random Forest with sum of multi-granularity word2vec embeddings (multi-granularity Word2vec+): four channels of embedding vectors of ICD codes are aggregated and then concatenated with the demographic features to create the patients representation.

CNN with one random initialization (one hot, rand): Each ICD code in the Specific level is randomly initialized and then modified during training.

CNN with dynamic single channel word2vec embeddings (single channel Word2vec): Each ICD code in the Specific level is initialized with pre-trained embedding vector and fine-tuned during training of the CNN model. Only one channel w.r.t the embedding of the Specific level ICD codes is used.

CNN with dynamic multi-granularity word2vec embeddings (multi-granularity Word2vec): ICD codes are initialized with pre-trained embedding vectors and fine-tuned. There are four channels of embedding vectors of ICD codes in four levels, w.r.t the embedding of the Category level, the Suborder level, the Detailed category level and the Specific level.

The linear regression and random forest model are performed using Scikit-learn 0.18.2 [29] and the number of trees for random forest is 10. The CNN models are implemented with Keras 2.0.6 [30]. All tasks are executed on the machine equipped with 8 NVIDIA Tesla K80 GPU cards. During training, we apply RMSprop algorithm [31] of batch size 1024 for stochastic optimization, with the initial learning rate set to 5e-3. And the max number iteration equals to 100. Besides, we also apply the learning rate decay schedule and the early stopping strategy to accelerate the convergence of training.

IV. Results

A. Prediction of Outcome

Table I shows results on the two prediction tasks, i.e. prediction of total hospital costs and length of stay. We can draw some conclusions:

TABLE I: The performance comparison of prediction of total hospital costs and length of stay. The value means R-squared value, 'Extremely Rare', 'Rare', 'Genearal' are corresponding to testing results got from 'extremely rare dataset', 'rare dataset' and 'general dataset', respectively. '–' means prediction task with rare dataset selection can't be carried out as the model can't tackle too high dimensional and sparse one-hot vectors. LR: Linear regression; RF: random forest; CNN: proposed convolutional neural network.

Regressor	Model	Total Hospital Costs			Length of Stay		
		Extremely Rare	Rare	General	Extremely Rare	Rare	General
LR	one-hot+	–	–	0.7316	–	–	0.4905
RF	one-hot+	–	–	0.7318	–	–	0.4548
	single channel Word2vec+	0.0252	0.2964	0.5599	0.0281	0.0186	0.3796
	multi-granularity Word2vec+	0.3681	0.5090	0.6788	0.1557	0.2276	0.4697
CNN	one-hot, rand	0.4534	0.6054	0.7660	0.2656	0.3237	0.5245
	single channel Word2vec	0.3794	0.5914	0.7648	0.2048	0.2840	0.5235
	multi-granularity Word2vec	**0.5085**	**0.6683**	**0.7876**	**0.2683**	**0.3491**	**0.5618**

Firstly, the proposed CNN model based on multi-granularity word embedding consistently performs best in both two applications, even the worst CNN model yields better performance than traditional regression model with one-hot representation. This reflects that the CNN is capable of extracting complex features and increases the predictive power. The advantage of embedding method that captures the semantic relationships between the medical concepts over the one-hot encoder contributes to more accurate predictive model. The MG-CNN model also has the best performance in the task of LOS prediction. There are some relations between the health condition and the length of stay. However, it is more difficult to predict the length of stay based on the medical codes and demographic characteristics. More information should be incorporated to improve the performance.

Secondly, the multi-granularity word embedding model outperforms the single channel model. In the total hospital costs prediction application, the overall improvement of CNN model with multi-granularity word embedding compared with the baseline model is around 4% when testing on the general dataset. The random forest model with multi-granularity word embedding model is 12% more accurate than the single channel word embedding. It proves that the multi-granularity word embedding can extract the essential and inner characteristics of the medical concepts and improve the prediction accuracy.

Thirdly, it is worth mentioning that the multi-granularity word embedding has great advantage in predicting the total hospital costs for the patients with rare medical events. For the rare dataset and the extremely rare dataset in which patients have many rare ICD codes, the prediction is much harder. The R-squared values of the best CNN model on the two dataset are just 0.5085 and 0.6683, respectively. Even the one-hot representation with traditional models cant tackle this prediction task with limited computational resource. Better yet, the improvement of the multi-granularity word embedding on these special datasets is much more obvious, which is around 8%-30% in term of R-squared value. This demonstrates our multi-granularity embedding method can handle the high dimensional and sparse data and improves the quality of embedding for the rare word.

Last but not least, the dynamic fine-tuning of embedding vector during the training of the CNN model results in better and more specific representation of the medical concept. Our MG-CNN does not require any pre-defined features, and it can adaptively learn high-quality features from large-scale dataset and use them for accurate prediction.

B. *Inspection of Embedding Concepts*

We inspect the medical concept vectors to understand the advantage of multi-granularity embedding. Fig. 6 shows the learned diagnosis vectors and procedures vectors plotted in a 2-D space, using t-SNE to reduce the dimension from 600 to 2. This is a technique for dimensionality reduction and particularly well suited for the visualization of high-dimensional datasets.

We observe that the diagnosis codes are generally well grouped by their corresponding categories or code initials. At the bottom of Fig. 6, we zoom in on three typical local regions. The left subfigure depicts diagnoses closely related to pregnancy, childbirth and the puerperium, including O20.001, O29.9x109, and other codes beginning with O; single stillbirth (Z37.102) and some related surgical operations like 69.5900. The middle subfigure depicts some closely related diagnoses of the blood and blood-forming organs, including D12.001 and other codes beginning with D; some diagnoses of intestines, including K62.602 and other codes beginning with K, and surgical operation 45.2400 which describes inspection of lower intestinal tract via natural or artificial opening endoscopy. The right subfigure shows some closely related diseases of the eye and adnexa, including H35.501 and other codes beginning with H, and some other diseases in different categories, but still related to eye and adnexa, such as Q14.302 whose description is congenital malformation of posterior segment of eye and Z97.001 described as having the presence of artificial eye (globe). From these three zoomed local regions, we confirm that our embedding vectors performed very well in naturally grouping closely related diseases and surgical operations.

Table III and Table II show the nearest neighbors of medical codes, including diagnostic code and procedure code. With

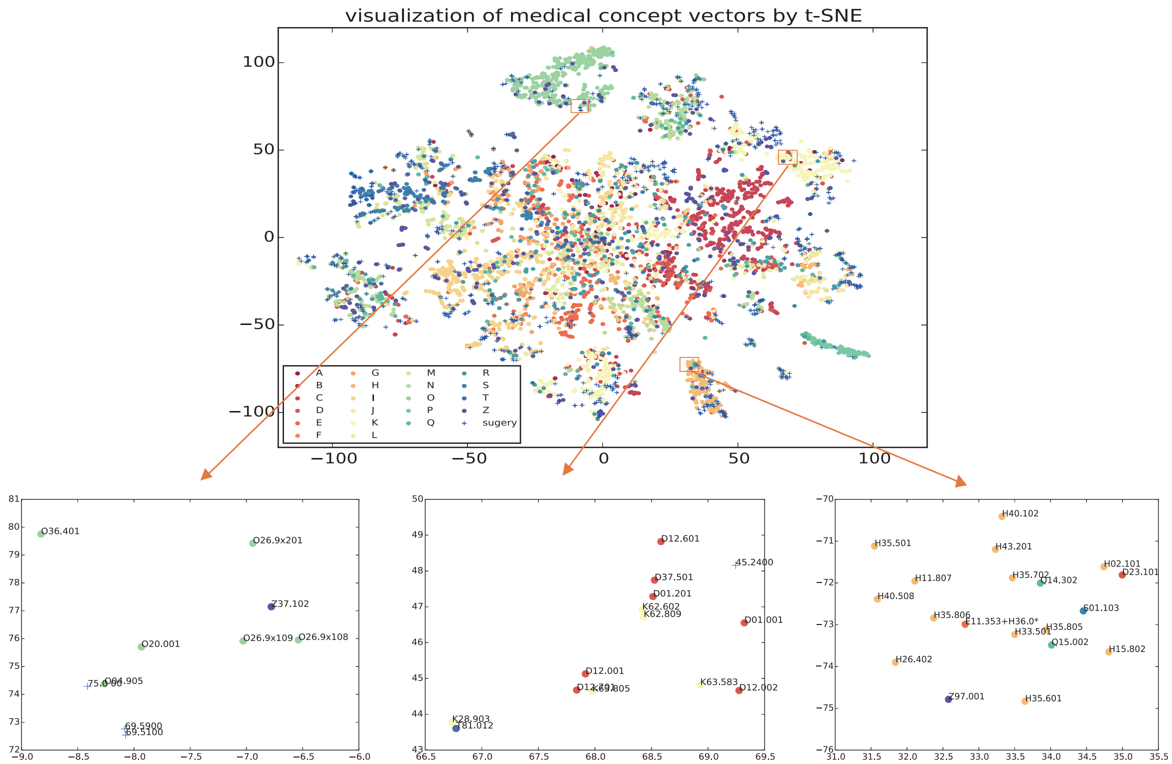

Fig. 6: Visualization of embedding concepts by t-SNE. One colored circle marker represents one diagnosis, or ICD-10 code, while one plus marker represents one procedure code, or ICD-9 code. According to the hierarchical grouping of ICD-10 codes, we divide circles into different subsets with different colors corresponding to the code initials in ‘A’ to ‘Z’.

the embedding vectors, we can find related diseases given a querying diagnostic code, and recommend relevant procedures to treat certain diseases. The similarity is evaluated by cosine distance of two medical concept embedding vectors. Thus, such embedding of vectors provides insights to practitioners to discover informative knowledge to inform clinical decisions.

TABLE II: The neighborhood of ICD-9 code 15.22 in the Detailed Category level which refers to Shortening procedure on one extraocular muscle. The top 5 neighbors are selected with filtering the duplicates.

ICD-9 code 15.22		
Name	Code	similarity
Recession of one extraocular muscle	15.11	0.9667
Operations on two or more extraocular muscles involving temporary detachment from globe, one or both eyes	15.30	0.9171
Astigmatism	H52.2	0.9134
Convergent concomitant strabismus	H50.0	0.8996
Vertical strabismus	H50.2	0.8809

V. Discussion

Some diagnoses are the natural complications of pre-existing conditions. In these cases, the underlying condition is the main diagnosis. We emphasize all the main diagnosis to further improve the predictive power. Each patient has one main diagnosis that indicates the disease condition at that time. There are 7936 main diagnosis codes in total. We concatenated the word2vec embedding vector of the primary ICD-10 diagnosis code with the demographic features and this improves the prediction accuracy around 1% in both two prediction tasks.

TABLE III: The neighborhood of ICD-10 code H50.301 in the Specific level whose description is intermittent exotropia. We display the top 5 neighbors, filtering the duplicates. The similarity is evaluated by cosine distance of two embedding vectors.

ICD-10 code H50.301		
Name	Code	similarity
Concomitant esotropia	H50.002	0.9557
Shortening procedure on one extraocular muscle	15.2200	0.9494
Hypermetropia	H52.001	0.9461
Superior oblique ophthalmoplegia	H49.804	0.9428
Esotropia	H50.004	0.9349

Certainly, our model can be further improved in some aspects. First, we can incorporate measurements to access the disease severity such as the injury Severity Score (ISS) or ICD-9 based Injury Severity Score (ICISS) which are measures of the quality of medical care. In addition, more strategies like attention mechanism can be integrated to build more interpretable deep learning model that can be more understandable. Eventually, with the explosive growth of clinical data, such model will be broadly applicable and provide much knowledge for clinical research.

VI. Conclusion

We introduce an effective multi-granularity embedding method for medical concepts representation and then propose a CNN model to estimate the impact of individual patient characteristics on the total hospital costs and LOS. Our MG-CNN model learns more accurate and reasonable representations to capture the latent relations between medical concepts by leveraging the prior medical ontology. Our method improves the quality of embedding for rare medical concepts and facilitates the measure of quality for the patients with low frequency event. In conclusion, MG-CNN model is able to discover potential knowledge from the clinical data to inform the clinical decisions.

Acknowledgment

This research is partially supported by the National Natural Science Foundation of China (61305066, 61561146396), the National Science Foundation of China (NSFC) under Grant 61561146396 and 71501109.

References

[1] J. Watts, “Chen zhu: from barefoot doctor to china’s minister of health,” *The Lancet*, vol. 372, no. 9648, p. 1455, 2008.

[2] M. of Health of the People’s Republic of China, “China public health statistical yearbook 2013,” 2013.

[3] “icd10,” https://www.cms.gov/Medicare/Coding/ICD10/index.html, accessed: 2017-10-07.

[4] “Unified medical language system (umls),” https://www.nlm.nih.gov/research/umls/, accessed: 2017-10-07.

[5] “Snomed ct,” https://www.nlm.nih.gov/healthit/snomedct/index.html, accessed: 2017-10-07.

[6] L. Jiang, H. M. Krumholz, X. Li, J. Li, and S. Hu, “Achieving best outcomes for patients with cardiovascular disease in china by enhancing the quality of medical care and establishing a learning health-care system,” *The Lancet*, vol. 386, no. 10002, pp. 1493–1505, 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[8] X. Min, Y. Zhou, S. Liu, and X. Bai, “Real-time object tracking via optimal feature subspace,” in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 421–425.

[9] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural probabilistic language model,” *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[10] R. Collobert and J. Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[11] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, “Deep patient: An unsupervised representation to predict the future of patients from the electronic health records,” *Scientific reports*, vol. 6, p. 26094, 2016.

[12] T. Tran, T. D. Nguyen, D. Phung, and S. Venkatesh, “Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm),” *Journal of biomedical informatics*, vol. 54, pp. 96–105, 2015.

[13] 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,” *arXiv preprint arXiv:1602.03686*, 2016.

[14] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: review, opportunities and challenges,” *Briefings in Bioinformatics*, p. bbx044, 2017.

[15] Y. Cheng, F. Wang, P. Zhang, and J. Hu, “Risk prediction with electronic health records: A deep learning approach,” in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 432–440.

[16] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, “Deepr: A convolutional net for medical records,” *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 22–30, 2017.

[17] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu, “Deep computational phenotyping,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 507–516.

[18] J. A. Minarro-Giménez, O. Marín-Alonso, and M. Samwald, “Exploring the application of deep learning techniques on medical text corpora.” *Studies in health technology and informatics*, vol. 205, pp. 584–588, 2013.

[19] L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, and P. Bruza, “Medical semantic similarity with a neural language model,” in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014, pp. 1819–1822.

[20] Y. Choi, C. Y.-I. Chiu, and D. Sontag, “Learning low-dimensional representations of medical concepts,” *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 41, 2016.

[21] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

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

[24] “rc020-icd-10Ł,” https://github.com/fengyujuan/MG-CNN, accessed: 2017-08-13.

[25] “rc022-icd-9,” https://github.com/fengyujuan/MG-CNN, accessed: 2017-08-13.

[26] W. Tingyan, Y. Ming, Y. Lan, N. Wenxin, and K. Dehua, “Code structure and information modeling for health intervention classifications,” *Journal of Tsinghua University (Science and Technology)*, vol. 65, no. 5, pp. 544–552, 2016.

[27] R. Rutledge, T. Osler, S. Emery, and S. Kromhout-Schiro, “The end of the injury severity score (iss) and the trauma and injury severity score (triss): Iciss, an international classification of diseases, ninth revision-based prediction tool, outperforms both iss and triss as predictors of trauma patient survival, hospital charges, and hospital length of stay,” *Journal of Trauma and Acute Care Surgery*, vol. 44, no. 1, pp. 41–49, 1998.

[28] “Hospital quality monitoring system,” https://www.hqms.org.cn, accessed: 2017-08-01.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] F. Chollet *et al.*, “Keras,” https://github.com/fchollet/keras, 2015.

[31] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.