

# A Comparison of Representation Learning Methods for Medical Concepts in MIMIC-IV

**Xuan Wu**

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China  
Email: wuxuan@mail.nju.edu.cn

**Yizheng Zhao**

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China  
Email: zhaoyz@nju.edu.cn

**Zhangdaihong Liu**

Oxford Suzhou Centre for Advanced Research, Suzhou 215123, China  
Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford OX1 2JD, UK  
Email: Jessie.Liu@oxford-oscar.cn  
Telephone Number: +86 0512-62869089

**Yang Yang**

School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, 200025, China  
Email: emma002@sjtu.edu.cn

**David A. Clifton**

Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford OX1 2JD, UK  
Email: david.clifton@eng.ox.ac.uk

## Abstract

**Objective:** To compare and release the diagnosis (ICD-10-CM), procedure (ICD-10-PCS), and medication (NDC) concept (code) embeddings trained by Latent Dirichlet Allocation (LDA), Word2Vec, GloVe, and BERT, for more efficient electronic health record (EHR) data analysis.

**Materials and Methods:** The embeddings were pre-trained by the four aforementioned models separately using the diagnosis, procedure, and medication information in MIMIC-IV. We interpreted the embeddings by visualizing them in 2D space and used the silhouette coefficient to assess the clustering ability of these embeddings. Furthermore, we evaluated the embeddings in three downstream tasks without fine-tuning: next visit diagnoses prediction, ICU patients mortality prediction, and medication recommendation.

**Results:** We found that embeddings pre-trained by GloVe have the best performance in the downstream tasks and the best interpretability for all diagnosis, procedure, and medication codes. In the next-visit diagnosis prediction, the accuracy of using GloVe embeddings was 12.2% higher than the baseline, which is the random generator. In the other two prediction tasks, GloVe improved the accuracy by 2%-3% over the baseline. LDA, Word2Vec, and BERT marginally improved the results over the baseline in most cases.

**Discussion and Conclusion:** GloVe shows superiority in mining diagnoses, procedures, and medications information of MIMIC-IV compared with LDA, Word2Vec, and BERT. Besides, we found that the granularity of training samples can affect the performance of models according to the downstream task and pre-train data.

**Keywords:** ICD-10-CM, ICD-10-PCS, NDC, code embedding, MIMIC-IV

## 1 Introduction

In the EHR systems, medical codes in diagnosis, procedure and medication contain rich medical information and abundant hidden knowledge about disease/therapeutic characteristics. Though the codes in different EHR system has distinct distribution due to different demographic of patient population, the hidden knowledge between these medical codes are universal. Existing EHR mining research [7, 31, 41] widely use pre-defined (also known as embedding) medical codes to capturing the hidden correlation between codes and enhancing various clinical tasks, such as diagnosis prediction and mortality prediction.

With the widespread deployment of EHR systems, clinical datasets such as MIMIC-IV [19] may store rich patient information while retaining their multimodality and complex structures. Well-structured code systems in EHR often record patient information such as diagnoses, procedures, and medications. A medical code is equivalent to a ‘word’ in natural language; all codes that occurred during a patient’s hospital visit naturally constitute one ‘sentence’; and a ‘document’ is equivalent to all hospital visits of a patient sorted by visiting time. The major inadequacy of this analogy is that each visit is a time-irrelevant set of codes. In other words, unlike natural language, the ‘words’ in each ‘sentence’ here do not appear in any particular order. As a result, order-insensitive sequence models based on the distributional hypothesis (LDA [2], Word2Vec [25], and GloVe [33]) are suitable for modelling medical codes. A more complex language model, BERT is also analyzed in this paper for comparison.

### 1.1 Word Embedding

In natural language processing (NLP), the semantic and syntactic features of words in unlabeled text data are expected to be fully captured in a low dimensional space through distributional semantics word embedding. Under distributional hypothesis [15], two semantically similar words are assumed to occur in a similar context, which means having similar word co-occurrence. Word2Vec [25] defines the context of the target word as the co-occurrence between target word and the words in the fixed-size window around it. As Word2Vec only considers the local co-occurrence, GloVe [33] utilizes global co-occurrence to capture explicitly defined sub-linear relationships. However, the drawback of these early word embeddings is that they represent all senses of a word into one embedding, while words are in multiple senses, in reality, e.g. “good” and “cheerful” are synonyms, but mean different in sentences “This restaurant has a good reputation.” and “A great many people would agree.”. Then many sense embedding works are proposed to give representations in different senses to a word in supervised or unsupervised ways. Some unsupervised sense embeddings works [28, 23] are based on LDA [2], a probabilistic topic model that can model the topic distribution over documents. In recent years, contextual embedding methods arises to give a solution that covers many aspects (e.g. sense representation, synonymy/antonymy, and hypernymy/hyponymy) [37]. The successful contextual embedding method BERT [10] learns contextual word embeddings based on bidirectional transformer architecture with the masked language model (MLM).

### 1.2 Medical Codes Representation Learning

Since the arising of deep learning, representation learning of medical codes becomes crucial for clinical prediction. Early works learn visit/patient representations [6, 6, 8, 29, 27, 4] utilizing deep learning methods such as recurrent neural networks, convolutional neural networks or stacked denoising autoencoders on longitudinal EHR data. These methods demonstrates the effectiveness of learning distributional representation for clinical prediction, but the representations are prediction-task-guided. For adaptation of different patient distribution in different EHR dataset, many works turn to utilizing language model to embed medical concepts. GRAM [7] learns ICD-9 disease code embeddings using GloVe and

shows a considerable improvement in sequential diagnosis predictions; G-BERT [38] pre-trains on ICD-9 disease codes for diagnoses and ATC codes for medications together with a modified MLM objective where the loss combines these two modalities of codes; BEHRT [22] embeds caliber codes for diagnoses with only MLM pre-training objective and achieved better results than visit and patient representation works [8, 27] in next visit prediction; Med-BERT [34] embeds ICD-9 and ICD-10 codes for diagnoses to predict the prolonged length of stay in hospital task with MLM pre-training. Recently, some works investigated the performance of other unsupervised NLP models in embedding medical concepts. Finch et al. [13] shows the clustering capability of ICD-10 disease codes with Word2Vec embeddings. Nuria et al. compared LDA and PLDA (supervised version of LDA) [21] ICD-10 in Osa dataset and ICD-9 in MIMIC-III dataset. However, for a variety of reasons, most of these works are unable to publish the pre-trained code embeddings. And as far as we know, there is no such comparison between these word embedding methods in longitudinal EHR data. The contribution of this work is twofold: (1) it gives an empirical comparison of four popular code embedding pre-training methods which can be considered as guidance for EHR researchers; (2) we are the first to publish the publicly accessible pre-trained embeddings<sup>1</sup> for ICD-10 codes using BERT, GloVe, Word2Vec, and LDA based on MIMIC-IV data, which can be used as initialization for any ICD-10 related tasks.

## 2 Method

### 2.1 Comparison Framework

To compare the pre-trained embeddings, LDA, Word2Vec, GloVe, and BERT were assessed in two ways: embedding interpretability and performance in three downstream tasks.

We set our pre-training to be downstream task-independent in order to make the application of the pre-trained embeddings as wide as possible. We note that the pre-training methods are unsupervised/self-supervised learning: the embeddings are learned from the intrinsic information in the data, such as the co-occurrence or the context information, and are not further tuned with the down-stream tasks.

The three modalities of medical concept, ICD-10-CM (Clinical Modification of diagnoses), ICD-10-PCS (Procedure Coding System), and NDC (National Drug Code), may have different compatibilities with LDA, Word2Vec, GloVe and BERT, therefore, were individually trained. Then the medical concept embeddings were fixed (without fine-tuning) and fed into downstream tasks as the input features of the concepts.

We extracted the diagnosis, procedure and medication information from the ‘diagnoses\_icd’, ‘procedures\_icd’ and ‘prescriptions’ tables under the ‘hosp’ module from the MIMIC-IV (version 1.0) website (<https://physionet.org/content/mimiciv/1.0/>), respectively. In particular, we focused on the ICD-10 codes for diagnosis and procedure and NDC codes for prescription. The statistical information of the data is depicted in Table. 1. All of the data were processed into two levels of granularities: each patient as a sample and each visit as a sample, as depicted in Figure. 1. When using a patient as a sample, we concatenated all visits of a patient ordered by time to form a ‘sentence’; when a visit serves as a sample, each ‘sentence’ is simply the collection of all codes that occurred during a visit. Notably, the codes within a visit are not ordered by time since this information is not provided in MIMIC. As far as we know, the choice of sample granularity has not been studied under the representation learning of medical concepts.

<sup>1</sup> <https://bit.ly/3ONj9Su>

Table 1. Information on diagnoses, procedures, and medications data. Avg. # of visits is on patients and Avg. # of codes is on visits.

Diagnoses				Procedures				Medications			
# concepts	# patients	Avg. # of visits	Avg. # of codes	# concepts	# patients	Avg. # of visits	Avg. # of codes	# concepts	# patients	Avg. # of visits	Avg. # of codes
17658	107704	2.49	11.79 ± 7.90	10429	61414	2.90	2.75 ± 2.49	8357	232064	2.13	21.21 ± 14.59

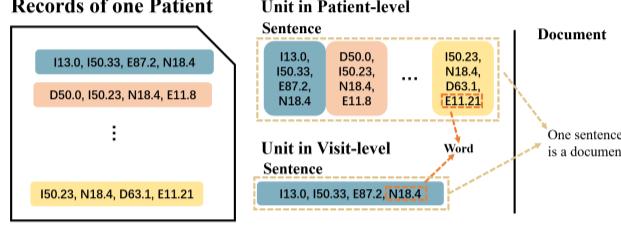


Figure 1: Data format for pre-training in the context of diagnoses data. Blocks with different colors are different visits. The same settings were used for procedures and medication data. A word in diagnoses data is an ICD-10-CM code, while for procedures and medications data is an ICD-10-PCS code, and an NDC code, respectively.

## 2.2 Models

### 2.2.1 Notation Definition

Let the sets of codes in diagnoses, procedures, and medications data be  $\mathcal{C}_D$ ,  $\mathcal{C}_P$ , and  $\mathcal{C}_M$ , respectively. We write  $\mathcal{C}$  as the set of codes when there is no need to specify the modality.  $\mathcal{C}$  is the code vocabulary. A sentence at visit-level is a list constructed by  $c \in \mathcal{C}$  of the form  $s_{ij} = [c_1, c_2, \dots, c_{v_{ij}}]$ , where  $v_{ij}$  is the amount of codes in the  $j$ th visit of the  $i$ th patient. While a sentence at patient-level is in the form  $s_i = [c_1, \dots, c_{v_1}, c_{v_1+1}, \dots, c_{\sum_{k=1}^j v_{ik}}, c_{\sum_{k=1}^j v_{ik}+1}, \dots, c_{\sum_{k=1}^{V_i} v_{ik}}]$ , where  $V_i$  is the amount of visits of the  $i$ th patient. The pre-training data consist of a document, which can be defined as  $\mathcal{D}$ . At visit-level,  $\mathcal{D} = [s_{11}, \dots, s_{1v_1}, \dots, s_{p1}, \dots, s_{pv_p}]$ , while at patient-level,  $\mathcal{D} = [s_1, s_2, \dots, s_p]$ , where  $p$  indicates the amount of patient.

### 2.2.2 LDA

Based on De Finetti's classic representation theorem [9], LDA mixture models capture the exchangeability of both words and documents. LDA assumes that each document is a mixture over latent topics, where each topic is characterized by a distribution over words. A topic distribution  $\theta_i$  for document  $\mathcal{D}_i$  is sampled from a Dirichlet allocation  $Dir(\alpha)$ . The topic  $z_{i,j}$  for the  $j$ th word in  $\mathcal{D}_i$  is sampled from the multinomial distribution  $Multinomial(\theta_i)$ . Sample  $\phi_{z_{i,j}}$  of the  $j$ th word  $c_{i,j}$  is from the dirichlet distribution  $Dir(\beta)$ . The word  $c_{i,j}$  then is modelled by the multinomial distribution  $Multinomial(\phi_{z_{i,j}})$ . Considering the successful application of LDA in learning phenotypes compared to other topic models, we assume that each dimension of an embedding is a latent topic, and the latent topics can be learned by LDA, so we used the word-topic matrix as the representation of the medical concepts. In this case, the embedding for each word is a vector of the probability in each topic, where the number of topics is the vector dimension. Then, the distribution is learned through variational inference. We implemented LDA using Python library *GenSim* [35], and the inference method is online variational Bayes [17].

### 2.2.3 Word2Vec

Word2Vec can make use of either continuous bag-of-words (CBOW) [25] or skip-gram [26]. The assumptions are different between CBOW and skip-gram. Skip-gram model assumes that a word can be used to generate its surrounding words, while CBOW assumes that the target word can be

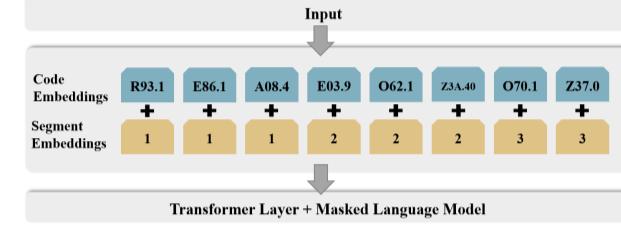


Figure 2: BERT model structure in the context of diagnoses data.

generated by its surrounding words. The surrounding words are the words around the target word within a fixed window size. We adopted the skip-gram version of Word2Vec in this work, as our early experiment showed that skip-gram has better performance in this case.

Given the dictionary  $\mathcal{C}$ , the training target of skip-gram is to maximize the average log probability:

$$\frac{1}{|\mathcal{C}|} \sum_{c=1}^{|\mathcal{C}|} \sum_{-w \leq j \leq w, j \neq 0} \log p(c_{t+j} | c_t) \quad (1)$$

The conditional probability  $p(c_{t+j} | c_t)$  is the probability of the presence of  $c_{t+j}$  given the context word  $c_t$  and is calculated as:

$$p(c_{t+j} | c_t) = \frac{\exp(\mathbf{E}_{ct+j}^{\text{in}} \top \mathbf{E}_{ct}^{\text{out}})}{\sum_{c \in \mathcal{C}} \exp(\mathbf{E}_c^{\text{out}} \top \mathbf{E}_{ct}^{\text{in}})} \quad (2)$$

where  $\mathbf{E}^{\text{in}}$  and  $\mathbf{E}^{\text{out}}$  are the 'input' and 'output' embeddings of codes. The final embedding for each code is the average between  $\mathbf{E}^{\text{in}}$  and  $\mathbf{E}^{\text{out}}$ .

We used the implementation of Word2Vec provided by Python library *GenSim* with hierarchical softmax and set the window size as 2.

### 2.2.4 GloVe

While Word2Vec learns the embeddings by mining the local context information, GloVe was proposed to globally compute the word embedding based on the matrix of word-word co-occurrence counts. Instead of considering the local co-occurrence within a fixed-size window, GloVe computes the co-occurrence within a data unit. In our use case, the data unit is the lists of clinical codes at visit/patient level. The distribution of code co-occurrences in GloVe is modelled by a power-law function of the frequency of each code pair. Based on the observation of the co-occurrence probabilities, GloVe models the ratio of co-occurrence probabilities.

### 2.2.5 BERT

BERT [10] uses a masked language model (MLM) to enable the pre-training of deep bidirectional embeddings, which can give both sentence embeddings and word embeddings. We take diagnosis data as an example and show the structure of BERT in Figure 2. The pre-training objective of MLM is to predict the randomly replace tokens into '[MASK]' tokens in the sentence with cross-entropy loss. In this way, code embeddings try to learn through their context. We pre-trained the MLM with 15% words masked in a sentence. We only used '[MASK]' tokens but not '[CLS]' and '[SEP]' tokens since only MLM was considered here.

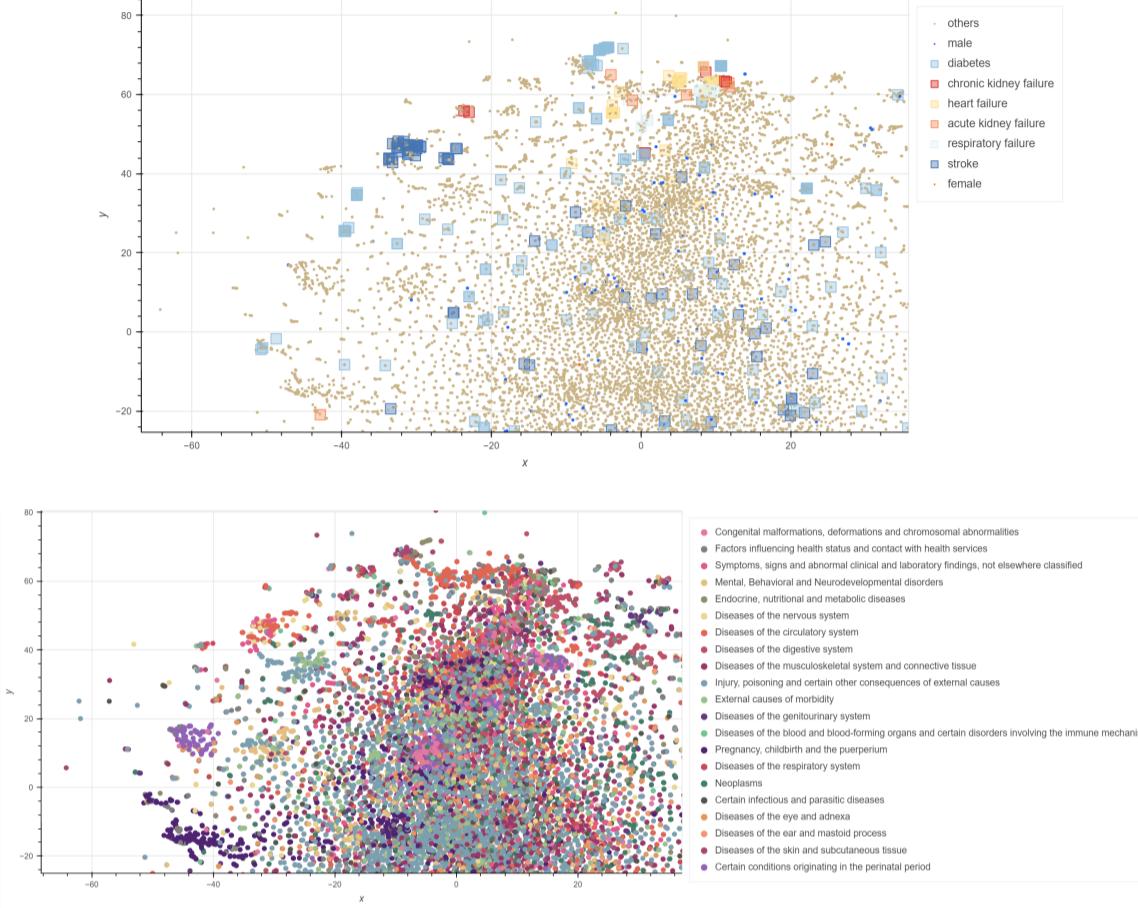


Figure 3: Visualization of GloVe pre-trained ICD-10-CM code embeddings reduced into 2D space by t-SNE, zoomed-in views. The points in these two subfigures are the same, but grouped by different categories. The points upper subfigure are classified into six typical diseases; the lower subfigure is classified according the top level of ICD-10 taxonomy.

#### 2.2.6 Settings

The model architectures are fixed for the three data modalities, diagnosis, procedure, and medication. To pre-train BERT, we only adopted the patient unit, as the visit unit is too short to show the advantage of BERT. For the other three methods, both of the data granularities were tested. The dimension of the embeddings was fixed as 128 for all models. In the pre-training, the optimizer Adam was employed. The learning rate was tuned as 5e-5, 1e-2, and 2.5e-2 for BERT, GloVe, and Word2Vec respectively. BERT was only pre-trained with MLM, following the setting of Med-BERT [34]. For Word2Vec, we set the window size as 2 with hierarchical softmax. The hyper-parameters in GloVe were set the same as the recommended settings in [33]. For LDA, the dimension of the embeddings was the number of topics.

### 3 Embedding Interpretability

To evaluate the interpretability of the embeddings learned by these models, we used t-SNE [40] to visualize the embeddings in 2D space, and we labelled points according to different standards. For ICD-10-CM codes, we classified points into six typical diseases as specified in Table 4. While the coding system of ICD-10-CM codes also provides a standard for classifying concepts, so we divided these codes into different categories

according to the coding system<sup>2</sup>. For ICD-10-CM codes, we visualize the embeddings with these two grouping standards and show them in Figure 3.

Each ICD-10-PCS code consists of seven characters<sup>3</sup>. The first character is the ‘section’, and the second through seventh characters mean different things in each section. For ICD-10-PCS, we used the 1st, 2nd, and 3rd characters to classify embeddings, respectively. For medication, we mapped the NDC codes into the Anatomical Therapeutic Chemical (ATC) code system [30], which is a hierarchical classification system that assigns a unique code to each medicine according to the organ system the medicine works on and how it works. We used the top-level ATC as the cluster labels which denote the main anatomical or pharmacological groups.

More visualizations can be seen in our repository. For diagnoses data, GloVe embeddings did the best in clustering diseases in kidney failure, stroke, and diabetes. LDA and BERT could also cluster chronic kidney failure well. All embeddings trained by the four methods cannot separate between male diseases and female diseases well. For procedure and medication data, embeddings pre-trained by GloVe also gave the best visualization.

To further show the clustering ability of the  $4 \times 3$  sets of embeddings, we used the clustering evaluation metric Silhouette coefficient [36].

<sup>2</sup> <https://www.icd10data.com/ICD10CM/Codes>

<sup>3</sup> <https://www.cms.gov/medicare/icd-10/2021-icd-10-pcs>

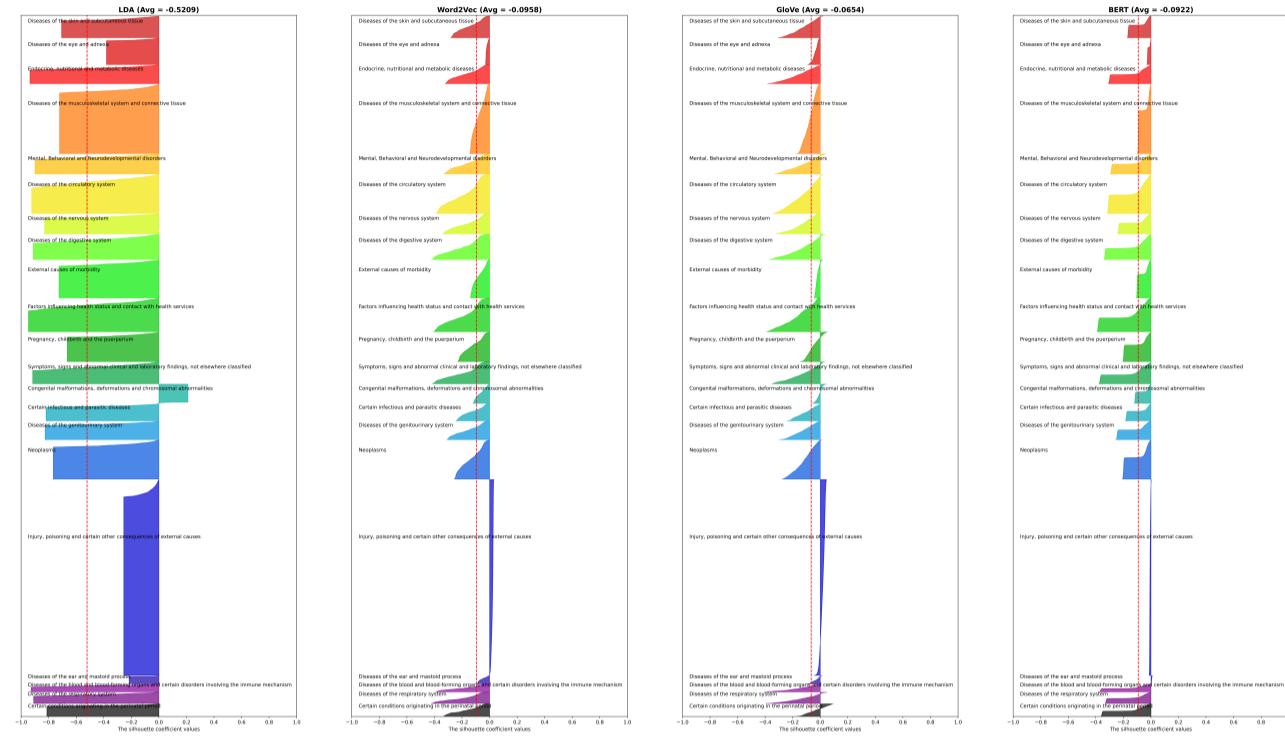


Figure 4: The silhouette coefficient value in clustering ICD-10-CM embeddings. The red dashed line denotes the average silhouette coefficient value.

The Silhouette coefficient is a value in the range of [-1, 1] measuring how similar points are within a cluster, where a larger value indicates a tighter cluster. Given embeddings and labels, the Silhouette coefficient for an embedding is  $\frac{b-a}{\max(a,b)}$ , where  $a$  is the mean distance between the embedding and all other embeddings in the same cluster, and  $b$  is the distance between the embedding and the nearest cluster that the embedding is not a part of. It is worth noting that all of these four models are not for clustering, so the application of Silhouette coefficient is mainly for the interpretability purpose. We computed the Silhouette coefficient in the original data space (128-d), and used one of the aforementioned labelling systems as clustering labels.

For ICD-10-CM embeddings, the visualization of Silhouette coefficients is shown in Figure. 4 with the top-level categories of ICD-10-CM. LDA clustered ‘Congenital malformations, deformations, and chromosomal abnormalities’ class well, while GloVe clustered ‘Certain conditions originating in the perinatal period’ better.

For ICD-10-PCS embeddings, we used the 2nd character as cluster labels, with results shown in Figure. 5. For most classes, LDA did the worst, for some small clusters, e.g., ‘Breast’, ‘Ear, Nose, Sinus’, ‘Lymphatic System’, LDA could knit them tight.

For NDC embeddings, in Figure. 6, the ones learned by Word2Vec clustered ‘Antineoplastic and immunomodulating agents’ well.

Overall, comparing the average silhouette coefficient, embeddings learned by GloVe displayed better clustering ability than the ones learned by other models.

## 4 EVALUATION

### 4.1 Tasks

To further compare the performance of these embeddings, we evaluated them in three tasks: diagnoses prediction of multiple diseases in the next visit within one month, mortality prediction for intensive care unit (ICU) patients, and medication recommendation. We set the diagnoses

prediction as a six-class classification task using diagnoses embeddings, predicting chronic kidney failure, acute kidney failure, heart failure, respiratory failure, diabetes, and stroke for the patients’ next visit. The mortality prediction is a binary classification task predicting whether mortality happens to ICU patients within a month of the current visit using diagnoses/procedures/medication embeddings. The medication recommendation is a regression task that uses the diagnoses/procedures embeddings to predict the medications within the same visit.

Patients with less than 2 visits are omitted in the evaluation. The patients were split into train and test sets with a ratio of 70% and 30%. Then patient-level data were further unwrapped to visit-level since one patient may have multiple visits.

Each training input of these tasks is a set of codes. The pre-trained embeddings for these codes are the feature matrix. To combine these codes’ feature vectors, we use the average of the code embeddings as the feature vector of each input. Except for the four sets of pre-trained embeddings, we also use randomly initialized embeddings to serve as baselines in the three modalities of data. We used four machine learning models—CatBoost [11], Xgboost [5], Random Forest [3] and Multi-layer Perceptron (MLP) [16] for classification (/regression).

#### 4.1.1 Diagnoses Prediction

As we know, multiple organ failure is a severe, life-threatening condition, while stroke and diabetes are correlated to multiple organ failure. The diagnoses prediction in this paper is the multi-classification between multiple organ failure diseases, stroke, and diabetes. This task was conducted on a cohort where patients have more than one visit. We further selected visit pairs with intervals of less than one month, and each visit was labelled with one of the six disease classes mentioned above according to the diagnosis code. Visits with multiple types of disease were also removed, so to rule out multi-label cases. The ICD-10 codes of these diseases are specified in Table 4. Lastly, visit pairs  $(v_1, v_2)$  with different labels were remained. We used the average of code embeddings in  $v_1$  to predict the disease label of  $v_2$ .

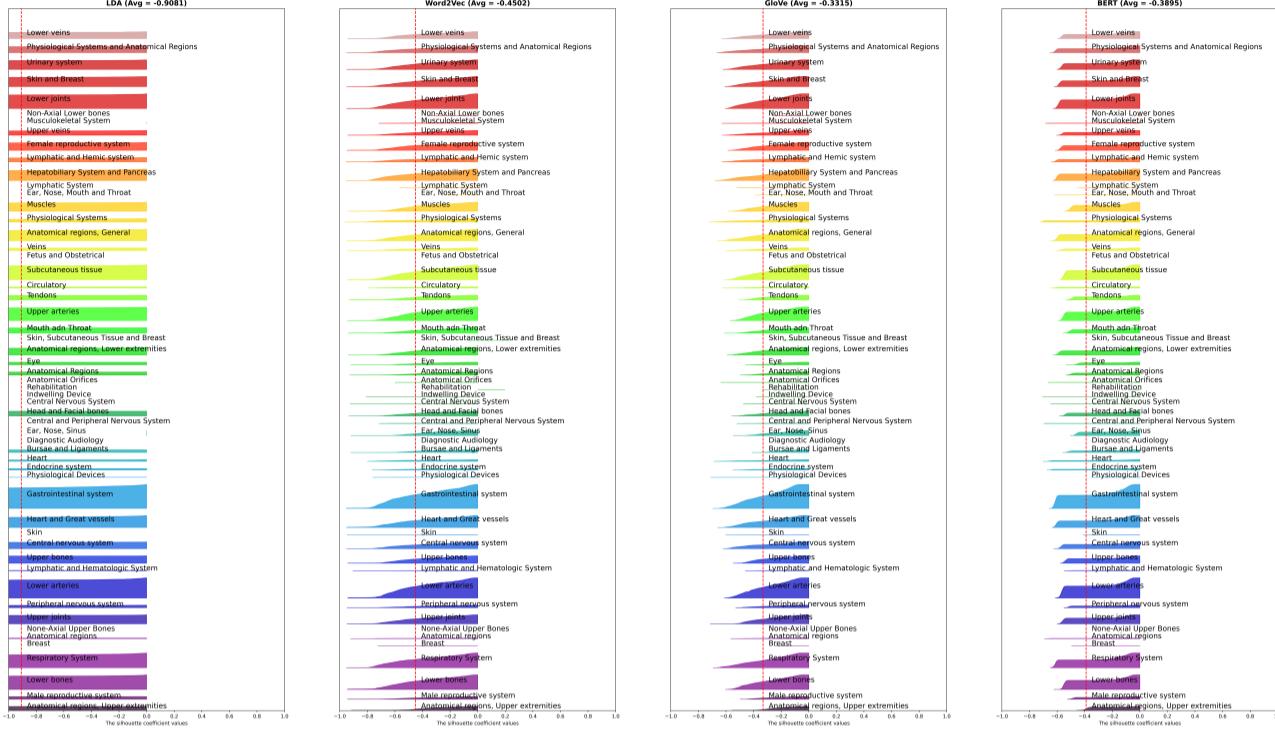


Figure 5: The silhouette coefficient value in clustering ICD-10-PCS embeddings.

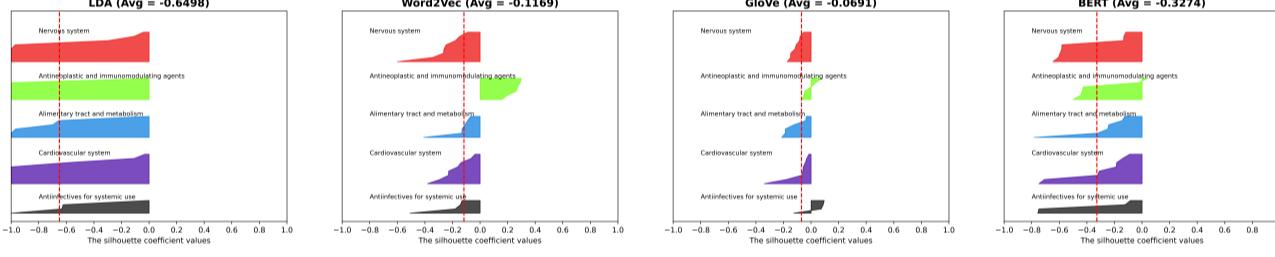


Figure 6: The silhouette coefficient in clustering NDC embeddings.

#### 4.1.2 Mortality Prediction

MIMIC-IV includes 21,622 ICU patients having diagnostic, procedures, and medication records. We collected the set of visits of these patients within a month since they entered the ICU, and unwrapped the set of visits into a set of medical codes. Each input is a set of medical codes of an ICU patient, and the task is to predict the label, which is 0/1 (denoting deceased or live within a month). As the positive samples, which are labelled as 1, are far more than the negative samples, we randomly down-sampled positive samples to the same size of the negative samples.

#### 4.1.3 Medication Recommendation

This task used the diagnostic/procedure information of a visit to predict the probability that each NDC would be in the prescription for this visit. To get the features of a visit, we firstly embedded the diagnostic/procedures codes in this visit and each prescription code for this visit using the pre-trained embedding from the same kind of model. Then we concatenated the averaged embedding of diagnostic/procedures embeddings and the embedding for one of the prescription codes as the feature of a sample with the training label as 1.

## 4.2 Results

The evaluation results are shown in the Table. 2 and Table. 3.

Table. 2 shows that all four methods have improved performance in diagnosis prediction compared with the baseline, randomly initialized embeddings. It indicates that the hidden information captured by these models has predictive powers for diagnoses. While in Table. 3, for procedures, the embeddings learned by LDA and Word2Vec did worse than the baseline. From Table. 2 and Table. 3, for the ICU patients mortality prediction, we can see that the diagnosis embeddings pre-trained by GloVe are more explicit in showing the patient's criticality in syndromes compared to procedures and medications data. While in the medication recommendation task, only embeddings learned by GloVe performed better than the randomly initialized inputs both for diagnoses and procedures. The distribution biases introduced by LDA could worsen the medication recommendation results. It may be caused by the poor interpretability of the LDA pre-trained embeddings of NDC codes.

Comparing the granularity of pre-training data, using pre-training data at the patient-level is better than using the visit-level for co-occurrence-based methods (GloVe, Word2Vec, and LDA) in the diagnoses prediction task. Pre-training with the visit-level data may lose the temporal relationship between diseases, especially for chronic diseases. However,

	Next Visit Diagnoses Prediction				Mortality Prediction				Medication Recommendation	
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Explained Variance	MSE
BERT	0.547	0.470	0.386	0.406	0.659	0.653	0.653	0.652	0.783	0.054
GloVe	0.632	0.598	0.520	0.548	0.693	0.695	0.695	0.693	<b>0.806</b>	<b>0.048</b>
GloVe (v)	<b>0.632</b>	<b>0.598</b>	<b>0.522</b>	<b>0.551</b>	<b>0.694</b>	<b>0.696</b>	<b>0.696</b>	<b>0.694</b>	0.805	0.049
Word2Vec	0.515	0.427	0.354	0.368	0.648	0.655	0.651	0.650	0.760	0.060
Word2Vec (v)	0.514	0.491	0.332	0.342	0.650	0.654	0.653	0.650	0.759	0.060
LDA	0.558	0.514	0.369	0.389	0.664	0.668	0.667	0.663	0.581	0.105
LDA (v)	0.548	0.506	0.362	0.388	0.658	0.664	0.662	0.658	0.607	0.098
random	0.510	0.395	0.307	0.311	0.657	0.660	0.660	0.657	0.759	0.060

Table 2. Evaluation results on diagnoses data. Model name with '(v)' means pre-training with the visit-level data. For classification tasks, we used accuracy, macro-precision, macro-recall, and macro-F1 as the metrics. To evaluate the regression task, we used explained variance and mean-squared error (MSE), where higher explained variance and lower MSE is better. We ran each experiment 5 times, and all of the standard deviations were less than  $2e - 2$ .

		Procedures								Medications							
		BERT	GloVe	GloVe (v)	Word-2Vec	Word2-Vec (v)	LDA	LDA (v)	random	BERT	GloVe	GloVe (v)	Word-2Vec	Word2-Vec (v)	LDA	LDA (v)	random
Mortality Prediction	Accuracy	0.668	<b>0.68</b>	0.678	0.647	0.648	0.659	0.639	0.659	0.621	0.648	<b>0.65</b>	0.623	0.62	0.632	0.636	0.62
	Precision	0.669	<b>0.681</b>	0.661	0.647	0.648	0.66	0.642	0.659	0.623	0.649	<b>0.652</b>	0.628	0.622	0.634	0.638	0.62
	Recall	0.668	<b>0.68</b>	0.678	0.647	0.648	0.659	0.64	0.659	0.623	0.649	<b>0.652</b>	0.622	0.622	0.634	0.638	0.618
	F1	0.668	<b>0.68</b>	0.678	0.647	0.648	0.658	0.639	0.659	0.62	0.647	<b>0.65</b>	0.618	0.62	0.631	0.636	0.617
Medication Recommendation	Explained Variance	0.780	<b>0.813</b>	0.803	0.787	0.788	0.757	0.76	0.787	/							
	MSE	0.055	<b>0.047</b>	0.049	0.053	0.053	0.061	0.06	0.053								

Table 3. Evaluation results on procedures and medications data.

for mortality prediction, as the critical factor is mainly determined by the fetal diagnoses [18] such as respiratory failure and sepsis, which is acute and often unexpected, there is no big difference between using the two kinds of data granularities. It is worth noting that GloVe showed the best performance for all three downstream tasks among all models considered. One of the reasons that GloVe outperformed BERT may be that BERT was not fine-tuned for these tasks, which we deliberately left out since we wanted to make the embeddings as generalized as possible. Embeddings learned by Word2Vec and LDA performed poorly on these tasks. Word2Vec does not use the global co-occurrence information, instead, it uses the context information within a fixed-size window; LDA is designed to do topic modelling. Nevertheless, embeddings learned by LDA have better performance than the ones learned by Word2Vec in the diagnoses prediction task and mortality prediction task.

## 5 DISCUSSION

Surprisingly, BERT did not do well in the evaluation. The reason may be that BERT is only pre-trained with MLM. Besides, the pre-training of BERT is sensitive to hyper-parameters and needs a large corpus, so the optimized pre-training method RoBERTa [24] can further be utilized for medical concepts representation learning.

With Word2Vec, the fixed window size introduces biases induced by the order of the NDC codes, which might be the main reason for the poor performances of the Word2Vec embeddings. In other words, the neighborhood of a medical code is restrained by the window size, leading to the omitting of codes that are out of the window but in the same visit.

For the tasks like clinical concept extraction in biomedical text, some published work has shown that BERT is better than GloVe and Word2Vec [39, 1, 32]. But in EHR data, Getzen et al. [14] also finds that BERT did not perform well when using static embeddings. Combining the results of this paper, we find that longitudinal EHR data is context-independent within a visit (i.e. the position of disease in the visit does not impact the semantics) but is context-dependent between visits. GloVe and Word2Vec word embeddings are context-independent, while BERT

generates different embeddings for a word to capture the context of the word, which is context-dependent. So BERT can introduce disturbance for medical code embeddings as there is no context but co-occurrence within a visit. And it is more suitable to use context-independent models e.g. GloVe and Word2Vec when using the visit as the sample for word embedding.

## 6 CONCLUSION

We have presented a new perspective on medical concepts representation learning methods comparison via visualizations and three downstream tasks. We considered three kinds of medical concepts: ICD-10-CM (diagnoses), ICD-10-PCS (procedures), and NDC (medications) and four pre-training methods — LDA, Word2Vec, GloVe, and BERT using the MIMIC-IV dataset. We compare the effectiveness of four methods in embedding three modalities of clinical codes. BERT pre-trained embeddings without fine-tuning cannot beat the embeddings learned by GloVe in this comparison framework. Furthermore, interestingly, embeddings learned by LDA are better than Word2Vec in the next visit diagnoses prediction task and mortality prediction task. Considering the visualizations and the results of the downstream tasks, pre-training with GloVe is the best in capturing the hidden semantics between medical codes in MIMIC-IV. Moreover, the granularity of pre-training samples should be considered according to different applications. In the mortality prediction, using visit-level data could be better, while in the medication recommendation, using patient-level samples may be a better choice.

## 7 ACKNOWLEDGMENT

## 8 SUPPLEMENTARY MATERIAL

The classified ICD-10-CM codes in Figure. 3 are specified according to Table. 4.

Diseases	ICD-10 codes	Num. of codes	Support
Chronic kidney failure	N18	7	26507
Acute kidney failure	N17,O904,T8612,N990	8	24737
Respiratory failure	J960,J962,J969,J9582,P285	12	11917
Heart failure	I50,I9713,I0981,I110,I130,I132	29	44832
Stroke	H341,I63,I64,I61,G45	101	5637
Diabetes	E10,E11,E12,E13,E14	158	63027

Table 4. Specified ICD-10 code prefix for six types diseases. We use the prefix to denote all of the codes with this prefix. The support is the occurrence amount of the codes in the specified class. The classification of codes refers to [20, 12].

## 9 DATA AVAILABILITY

The MIMIC-IV data underlying this article is available in <https://doi.org/10.13026/a3wn-hq05> with restricted-access.

## 10 FUNDING

### References

- [1]E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.
- [2]D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [3]L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [4]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*, pages 507–516, 2015.
- [5]T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [6]E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun. Multi-layer representation learning for medical concepts. In *proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1495–1504, 2016.
- [7]E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 787–795, 2017.
- [8]E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.
- [9]B. De Finetti. *Theory of probability: A critical introductory treatment*, volume 6. John Wiley & Sons, 2017.
- [10]J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [11]A. V. Dorogush, V. Ershov, and A. Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.
- [12]J. Dugan and J. Shubrook. International classification of diseases, 10th revision, coding for diabetes. *Clinical Diabetes*, 35(4):232–238, 2017.
- [13]A. Finch, A. Crowell, M. Bhatia, P. Parameshwarappa, Y.-C. Chang, J. Martinez, and M. Horberg. Exploiting hierarchy in medical concept embedding. *JAMIA open*, 4(1):ooab022, 2021.
- [14]E. Getzen, Y. Ruan, L. Ungar, and Q. Long. Mining for health: A comparison of word embedding methods for analysis of ehrs data. *medRxiv*, 2022.
- [15]Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.
- [16]G. E. Hinton. Connectionist learning procedures. In *Machine learning*, pages 555–610. Elsevier, 1990.
- [17]M. Hoffman, F. Bach, and D. Blei. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23, 2010.
- [18]J. C. Jentzer, S. van Diepen, D. H. Murphree, A. S. Ismail, M. T. Keegan, D. A. Morrow, G. W. Barsness, and N. S. Anavekar. Admission diagnosis and mortality risk prediction in a contemporary cardiac intensive care unit population. *American heart journal*, 224:57–64, 2020.
- [19]A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark IV. Mimic-iv (version 0.4). *PhysioNet*, 2020.
- [20]R. A. Kokotailo and M. D. Hill. Coding of stroke and stroke risk factors using international classification of diseases, revisions 9 and 10. *Stroke*, 36(8):1776–1781, 2005.
- [21]N. Lebeña, A. Blanco, A. Pérez, and A. Casillas. Preliminary exploration of topic modelling representations for electronic health records coding according to the international classification of diseases in spanish. *Expert Systems with Applications*, page 117303, 2022.
- [22]Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020.
- [23]Y. Liu, Z. Liu, T.-S. Chua, and M. Sun. Topical word embeddings. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.
- [24]Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [25]T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [26]T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.
- [27]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*, 6(1):1–10, 2016.
- [28]D. Q. Nguyen, A. Modi, S. Thater, M. Pinkal, et al. A mixture model for learning multi-sense word embeddings. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (\* SEM 2017)*, pages 121–127, 2017.
- [29]P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh. Deepr: a convolutional net for medical records. *IEEE journal of biomedical and health informatics*, 21(1):22–30, 2016.
- [30]W. H. Organization et al. The anatomical therapeutic chemical classification system with defined daily doses-atc/ddd. 2009.
- [31]X. Peng, G. Long, T. Shen, S. Wang, and J. Jiang. Sequential diagnosis prediction with transformer and ontological representation. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 489–498. IEEE, 2021.
- [32]Y. Peng, S. Yan, and Z. Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, 2019.
- [33]J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on*

- 
- empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [34]L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13, 2021.
- [35]R. Rehurek and P. Sojka. Gensim—python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2):2, 2011.
- [36]P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
- [37]E. Sezerer and S. Tekir. A survey on neural word embeddings. *arXiv preprint arXiv:2110.01804*, 2021.
- [38]J. Shang, T. Ma, C. Xiao, and J. Sun. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*, 2019.
- [39]Y. Si, J. Wang, H. Xu, and K. Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304, 2019.
- [40]J. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [41]K. Yu, M. Zhang, T. Cui, and M. Hauskrecht. Monitoring icu mortality risk with a long short-term memory recurrent neural network. In *PACIFIC SYMPOSIUM ON BIocomputING 2020*, pages 103–114. World Scientific, 2019.