Heterogenous Graph Embeddings of Electronic Health Records Improve Critical Care Disease Predictions

Tingyi Wanyan[1,2,3,4], Martin Kang[5,6], Marcus A Badgeley[5,7,8], Kipp W Johnson[2], Jessica K De Freitas[1,2], Fayzan F Chaudhry[1,2], Akhil Vaid[1,2], Shan Zhao[1], Riccardo Miotto[1,2], Girish N Nadkarni[1,2,9,10], Fei Wang[11], Justin Rousseau[12], Ariful Azad[4], Ying Ding[3,4,12], and Benjamin S Glicksberg[1,2,*]

[1] Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY
[2] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY
[3] Intelligent System Engineering, Indiana University, Bloomington, IN
[4] School Of Information, Uniersity of Texas Austin, TX
[5] nference, Cambridge, MA
[6] Department of Dermatology, Mayo Clinic, Rochester
[7] Department of Anesthesiology, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA
[8] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA
[9] Department of Medicine, Division of Nephrology, Icahn School of Medicine at Mount Sinai, New York, NY
[10] Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY
[11] Department of Health Policy and Research, Weill Cornell Medical School, Cornell University, New York, NY
[12] Dell Medical School, University of Texas, Austin, TX

Correspondence: Benjamin Glicksberg, benjamin.glicksberg@mssm.edu

Abstract. Electronic Health Record (EHR) data is a rich source for powerful biomedical discovery but it consists of a wide variety of data types that are traditionally difficult to model. Furthermore, many machine learning frameworks that utilize these data for predictive tasks do not fully leverage the inter-connectivity structure and therefore may not be fully optimized. In this work, we propose a relational, deep heterogeneous network learning method that operates on EHR data and addresses these limitations. In this model, we used three different node types: patient, lab, and diagnosis. We show that relational graph learning naturally encodes structured relationships in the EHR and outperforms traditional feed forward models in the prediction of thousands of diseases. We evaluated our model on the EHR data derived from MIMIC-III, a public critical care data set, and show that our model has improved prediction of numerous diagnosis.

Keywords: Electronic Health Records, · Heterogeneous Graph Learning · Skip-gram Model · Embeddings

1 Introduction

Electrical Health Records (EHRs) have rapidly emerged over the past 10 years as a powerful source for biomedical resource [11]. EHRs consist of clinical data from patient encounters with healthcare systems, which include demographic information, diagnoses, laboratory tests, medications, and clinical notes. EHR data have been used to develop machine learning (ML), and deep learning (DL), models for predicting diagnoses, mortality, length of hospital stay, and future illnesses. However, many of these ML-related solutions to clinical tasks consist of simple rule based models that, while possibly are easier to implement, often do not capture the complex patterns of the data. Some of these solutions are sufficient for certain clinical tasks, but for other tasks they are lacking. For example, one factor for determining priority for transplantation is a model for end-stage liver disease which includes only four variables and was trained on only 231 patients [8]. While there are a number of barriers that need to be overcome for DL to pervade healthcare operations, one particular hurdle is developing more suitable EHR representations for modeling.

Current EHR systems are constructed with numerous medical codes of different types to represent diverse data elements captured in clinical encounters. The performance of DL models on EHR could benefit from accurately capturing and modeling these heterogeneous data [2,10]. The most common approach to handle disparate data types is to treat each patient encounter as an unordered set of features, and concatenate these features together as the input to a DL system [10]. Such an approach is straightforward, intuitive, and easy to manipulate. However, this feature integration approach disregards the graphical structure and inner connectivity between medical concepts, such as physician's decision process [4]. More importantly, the lack of encoding patient to patient similarity makes this approach lose many information that pairwise patients could provide in various aspects, such as cohort analysis, disease sub-typing, diagnoses comparison, and treatment comparison [12]. Some recent graphical techniques emerges on modeling the connectivity nature, such as the graph model that captures the physician's treatment procedure [3, 4], and a temporary graph model [7] that captures the medical concepts inner connectivity patterns. But these techniques performs per-patient training, lack of considering the information provided from similar patient.

Furthermore, DL modeling is difficult because of issues of data quality [6] due to insufficient patient information and missing values among others. Besides that, data diversity and non-uniform length of time series within each patient also create issues for modeling. For example, patient encounter frequency varies in length, ranging from only one encounter to multiple readmissions. Also, length of stay could vary from a few hours to several months. The data sparsity along with data diversity create difficulties for deep learning models such as LSTM

system [2, 6], which requires abundant training data in order to reach good performance.

In this paper, we propose a Heterogeneous Graph learning Model (HGM) and apply it to EHR data. It contains various techniques and properties that attempts to overcome the aforementioned problems. Since it is a graph structure, this model could more naturally capture the inter connectivity between medical concepts. It also connects similar patients by their disease profiles, so that information from a similar patient could be leveraged for encoding in the target patient representation. The graph model learns representations by propagating information through the whole network, so when the data set is sparse, the embedding representation for each patient could be learned from the information traversing the whole network. This model, learning using Skip Gram With Negative Sampling strategy [9], is an efficient way of using all complex information available at hand. We show that with the relational heterogeneous graph learning, we can reach marginal improvement on diagnoses classification accuracy given patients' lab tests against traditional per-patient training strategy using shallow multilayer perceptron neural network.

2 Methodology

In this section, we introduce the theoretical construction of our heterogeneous EHR graph model. Please refer to the Supplementary Materials (https://github.com/Tingyiwanyan/mfgcn/tree/master/src/supp) for an in depth description of the preliminaries for these models.

2.1 Data set

For this work, we utilized EHR data from the critical care MIMIC-III de-identified data set. This data set is comprised of various elements relating to patients during their hospital care, such as demographics, lab test results, disease diagnoses, among others. We sampled the first 3801 patients in the dataset and collected all of their associated lab tests and diagnoses. These patients had received 447 unique lab tests and 2922 unique diagnoses. The limitation on the cohort sample size was due to the RAM required to load all of these data as a graph into memory.

2.2 Data Representation for Graphical Model

We create a graphical model of the EHR data by representing patients, labs, and diagnoses as nodes in a directed graph. Nodes are connected by edges, which come in two flavors and can be represented with the triples:

$$Lab \xrightarrow{testing} Patient : \{Lab, testing, Patient\}$$

$$Patient \xrightarrow{diagnosed} Diagnosis : \{Patient, diagnosed, Diagnosis\}$$

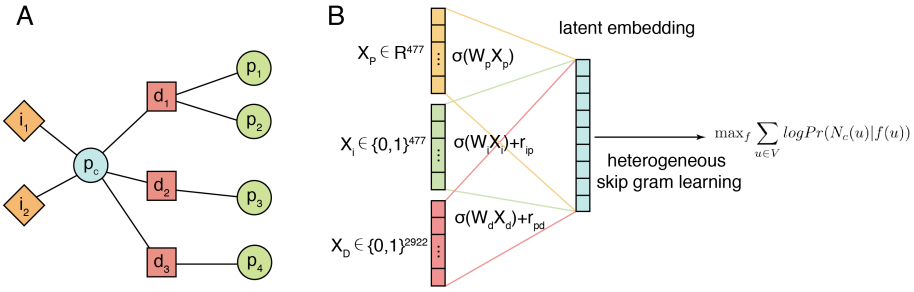

Fig. 1. Model schematics for representing EHR data in a heterogenous graphical model (A) and dense vectors (B). All graph nodes in (A) have a corresponding vector like those shown in (B). The vector representations can be projected into a shared space with the TransE method, and this projection optimized for retaining relations in the original data in the embedding via skip-gram optimization.

The initial *Patient* node representation is a vector $X_p \in R^{477}$ containing the measured values from lab tests. We initially represented labs and diagnoses as one-hot encodings: $X_l \in \{0,1\}^{477}$ and $X_d \in \{0,1\}^{2992}$.

With these two types of relationships T_E, we can construct the heterogeneous graph integrating the specified elements of the entire EHR data (Figure 1A). One *Patient* node could have connection with multiple *Diagnoses* node, and these *Diagnoses* nodes could link to other patients who have the same ICD code.

2.3 Embedding the HGM into a Latent Space

Nodes from a HGN can be embedded into a shared latent space using the TransE method (Figure 1B) [1]. This method uses a set of 1) projection matrices and 2) relation vectors. After initialization, projections and translations can be optimized end-to-end (see section 2.4).

HGM nodes X_p, X_l, X_d are projected into a shared latent space with with trainable projection matrices W_p, W_i, W_d using these nonlinear mappings:

$$c_p = \sigma(W_p \cdot X_p)$$
$$c_i = \sigma(W_i \cdot X_i)$$
$$c_d = \sigma(W_d \cdot X_d)$$

Where σ is a non-linear activation function and c_p, c_i, c_d are the latent representations of each type of node. Despite the EHR-space using different dimensions for different node types X_p, X_i, X_d, all nodes types are projected into the same latent space.

Then we apply translation operations to link these different types of nodes:

$$c_p = c_i + r_{ip}$$
$$c_d = c_p + r_{pd}$$

Where r_{ip} and r_{pd} are the relation vectors connecting patients to labs and diseases, respectively.

2.4 Optimizing the HGM Embedding

With the projection and translation operations we can convert different types of nodes into the same latent space. We then tune these parameterized transforms to increase the proximity between those embedding points whose corresponding graph nodes are often connected. Specifically, we apply Heterogeneous Skip-gram optimization using the optimization model [5]:

$$\max \sum_{u \in V} \sum_{t \in T_V} logPr(N_t(u)|f(u)) \tag{1}$$

Where $N_t(u)$ is the heterogeneous neighborhood vertices of center node u, and $t \in T_V$ is the node type. Here, we learn effective node embeddings by maximizing the probability of correctly predicting the a patient node's associated labs and diagnoses. The prediction probability is modeled as a softmax function:

$$Pr(c_t|f(u)) = \frac{e^{\vec{c}_t \cdot \vec{u}}}{Z_u} \tag{2}$$

Where $\vec{u}$ is the latent representation of patient u, $\vec{c}_t$ is the latent representation of lab and diagnosis neighbors of node of u, and $\vec{c}_t \cdot \vec{u}$ is the inner product of the two embedding vectors representing their similarity. Z_u is the normalization term $Z_u = \sum_{v \in V} e^{\vec{v}_t \cdot \vec{u}}$. Where Z_u integrate over all vertices. Therefore, equation 1 could be simplified to:

$$\mathcal{L}_s = -\sum_{t \in T} \sum_{u \in V} \Big[\sum_{c_t \in N_t(u)} \vec{c}_t \cdot \vec{u} - logZ_u \Big] \tag{3}$$

Numerical computation of Z_u is intractable for very huge graph with millions of nodes. So we adopt negative sampling strategy [9] to approximate the normalization factor, and the optimization function becomes:

$$\mathcal{L}_s = -\sum_{t \in T} \sum_{u \in V} \Big[\sum_{c_t \in N_t(u)} log\sigma(\vec{c}_t \cdot \vec{u}) + \sum_{j=1}^{K} E_{c_j \sim P_v(c_j)} log\sigma(-\vec{c}_j \cdot \vec{u}) \Big] \tag{4}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$, K is the number of negative samples. $P_v(c_j)$ is the negative sampling distribution. Equation 4 is the final objective function we are using for heterogeneous graph learning.

For training our Heterogeneous Graph Model (HGM), we perform heterogeneous neighborhood sampling by its one-hop connectivity, and pick *Patient* node as the center node, since it has one-hop connections to both *Diagnoses* and *Item_test* nodes. Specifically, for one training center *Patient* node, we uniformly sample 10 *Diagnoses* one-hop direct connected nodes, and 10 *Item_test* one-hop direct connected nodes. From these sampled 10 *Diagnoses* nodes, we sample

another 10 *Patient* nodes, each has connection with each of the 10 *Diagnoses* nodes. In this way, we connect the center patient node with its similar other *Patient* nodes by their common diagnoses. For negative sampling [9], we perform uniform sampling through all *Diagnoses* node and *Item_test* nodes that don't have one-hop connections with the center training patient node. Then we project these different nodes into same latent space through TransE model, after unifying the embeddings for different node types, each concept is representetd as a point in a Euclidean space. In this space we can measure the similarity between any two points by the angle between vectors between them and the origin.

2.5 Disease Prediction

For diagnosis prediction, we used the HGM embedding vectors to identify similar patients and diagnoses, and evaluate how this approach compares to a classical feed forward Neural Network approach. We record F1 score and AUC score as the evaluation metric for comparison.

We split patients into a group of 2,660 used to fit the MLP and HGM embedding and 1,141 used to evaluate disease prediction. For each patient, we computed the distance between a patient plus the diagnosis translation vector r_{pd} to all diseases.

The baseline MLP model is a feed-forward encoding-decoding neural network structure with a single hidden embedding layer whose dimensionality matches the embeddings produced by the HGM. The decoding part is a softmax layer for classifying correct diagnoses, so the MLP is trained in a supervised fashion.

3 Results

3.1 Embedded Representation

By learning a heterogenous graph embedding for each node type and then using transE to translate between type-specific embeddings, we generate dense vector representation in a space shared between all node types (Figure 3). There's a mixed cluster of all node types and several type-selective clusters.

Upon inspection, salient clusters of labs tests can be identified when the embeddings are projected into 2D TSNE space for visualization. The members of one cluster corresponded to routine comprehensive metabolic panels, while members of another cluster largely consisted of ventilator measurements.

3.2 Diagnosis Prediction Performance Comparison

The HGM outperformed the MLP on many diagnoses. When evaluating both models' diagnosis predictions across all common diagnnoses, the HGM has a higher performance than an MLP across all tested latent embedding dimensionalities (Table 1, Figure 2). Notably, the performance of HGM remained consistent with larger embeddings, while the performance of MLP degraded with larger embeddings.

Table 1. Diagnosis Classification Performance

Model	F1 score	AUC score
100 hidden latent embedding dimension		
MLP-Sigmoid	0.671	0.788
MLP-Tanh	0.517	0.778
MLP-Relu	0.483	0.765
HGM-Sigmoid	**0.739**	**0.834**
HGM-Tanh	0.727	0.829
HGM-Relu	0.713	0.839
200 hidden latent embedding dimension		
MLP-Sigmoid	0.625	0.766
MLP-Tanh	0.447	0.755
MLP-Relu	0.446	0.746
HGM-Sigmoid	**0.741**	**0.835**
HGM-Tanh	0.733	0.828
HGM-Relu	0.739	0.840
500 hidden latent embedding dimension		
MLP-Sigmoid	0.537	0.753
MLP-Tanh	0.377	0.724
MLP-Relu	0.419	0.734
HGM-Sigmoid	**0.751**	**0.834**
HGM-Tanh	0.735	0.829
HGM-Relu	0.743	0.842

The predictive performance of these models varied widely by disease, as shown in (Figure 3B). The performance of HGM was particularly strong with diseases that were more prevalent in the test set (see Table 2). We observed only one diagnosis, end stage renal diseases (ESRD), where MLP outperformed HGM (MLP F1: 0.606, HGM F1: 0.245) (Figure 3A).

For the diagnoses with at least 1 percent in prevalence, The median, 25th percentile, and 75th percentile of MLP predictive F1 scores are 0, 0, 0, respectively. The range of MLP F1 distribution is 0 to 0.606. For the same set of diagnoses, the median, 25th percentile, and 75th percentile of HGM F1 scores are 0.041, 0.024, 0.081, respectively, and the range of HGM F1 distribution is 0 to 0.562.

4 Discussion

In this work, we present HGM embeddings as a way to naturally represent EHR data relations with dense vectors and an embedding space containing all the node types in the original graph. By measuring distances between patient and disease concepts in this embedding space, we were able to predict which diagnoses a group of hold-out patients had with better performance than a supervised model trained specifically to predict patients' diagnoses from their labs.

Table 2. Prediction Performance on Most Observed Diagnoses

Diagnoses	**HGM-sigmoid** F1 score	**MLP-sigmoid** F1 score
Congestive heart failure	0.562	0.406
Unspecified essential hypertension	0.512	0.435
Atrial fibrillation	0.447	0.423
Acute kidney failure	0.455	0.415
Coronary atherosclerosis	0.365	0.163
Other and unspecified hyperlipidemia	0.367	0.297
Acute respiratory failure	0.316	0.067
Esophageal reflux	0.311	0.041
Diabetes mellitus	0.297	0.192
Urinary tract infection	0.276	0.060

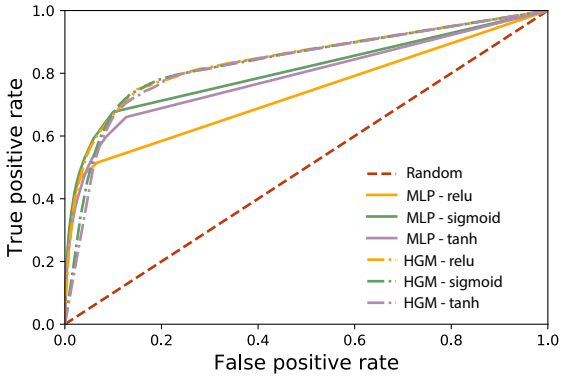

Fig. 2. Binary classification performance of HGM and MLP across common diseases. Each line represents the tradeoff of sensitivity and specificity for a classifier. HGM frameworks with larger embedding spaces perform better than the MLP models.

Averaged across all diseases, the HGM consistently outperformed the MLP across a range of activation functions and embedding dimension sizes. HGM and MLP had different trends as the dimensionality of the embedding increased. A larger embedding provides more complex representations, but is more likely to be overfit to training data. As the dimensionality of the embedding was increased, the MLP AUC decreased with a dosage effect observed across all activation functions. However, the HGM maintained a stable performance across all embedding capacities. This suggests that the embeddings learned by HGM are less susceptible to overfit training data.

MLP outperformed HGM on only one diagnosis, End Stage Renal Diseases (ESRD). This may be because the diagnosis of ESRD can be determined solely by a single lab test, estimated Glomerular Filtration Rate (eGFR). Thus, it is

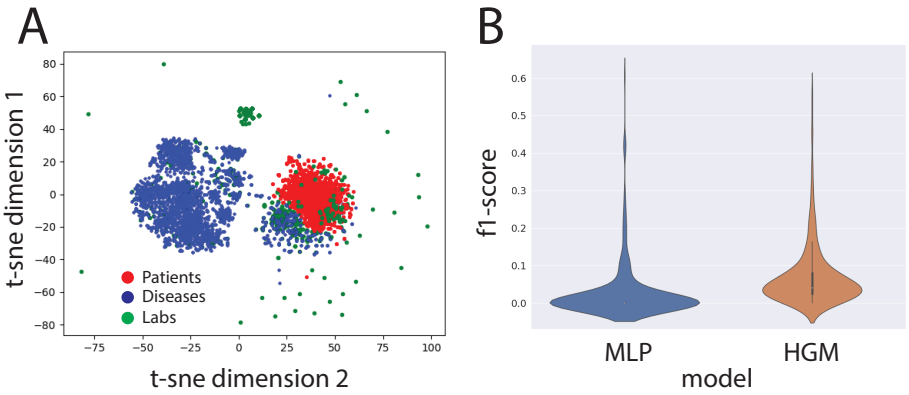

Fig. 3. A) R^2 view of the R^{500} embedding space shared by all data types. Each point represents a graph node, and that node's type is indicated with color. The dimensionality of the space is reduced for visual interpretation with tSNE. B) Distribution of F1 scores for common diagnoses using a HGM or MLP model. Diagnoses with at least 1% prevalence in the test set were included.

less likely that the prediction of ESRD will benefit from the graphical property of HGM.

One key feature of HGM is that the graphical structure of HGM explicitly declares and takes the sum of information from all patient nodes connecting a given pair of diagnosis and lab test. On the contrary, MLP flattens all features at the patient-level, and performs training on the per-patient basis, allowing only indirect connections between pairs of diagnoses and lab tests and relying only on network parameters to learn the underlying biomedical relationships.

The data set we used to fit HGMs allowed us to develop an EHR-knowledge graph across the compendium of care provided in Intensive care units. We found that our model was robust to overfitting but there may be bias in lab-disease relationships between patient populations or intensive care practices. Only the sickest patients are admitted to the ICU, so this model should be fine-tuned for other inpatient applications. Another consequence of only having ICU visits is that most people have only 1 or few ICU admissions, which is not suitable for time series models. Other studies applying graph theory to EHRs have been able to perform robust sequential diseases prediction that consistently outperforms non-graph models [3].

5 Conclusion

In this study, we apply a deep heterogeneous graph model(HGM) to learn the representation of EHR data. In the task of diagnosis prediction, HGM embeddings consistently outperformed non-graphical baseline models across diagnoses and appears less susceptible to overfitting of training data. Our findings suggest that HGM is a promising strategy to develop generalizable EHR-knowledge

graph. In the future, we expect to apply HGM to other clinically relevant tasks and assess performance across multi-insitutional datasets.

References

1. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in neural information processing systems. pp. 2787–2795 (2013)
2. Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F., Sun, J.: Doctor ai: Predicting clinical events via recurrent neural networks. In: Machine Learning for Healthcare Conference. pp. 301–318 (2016)
3. Choi, E., Xiao, C., Stewart, W., Sun, J.: Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In: Advances in neural information processing systems. pp. 4547–4557 (2018)
4. Choi, E., Xu, Z., Li, Y., Dusenberry, M.W., Flores, G., Xue, Y., Dai, A.M.: Graph convolutional transformer: Learning the graphical structure of electronic health records. arXiv preprint arXiv:1906.04716 (2019)
5. Dong, Y., Chawla, N.V., Swami, A.: metapath2vec: Scalable representation learning for heterogeneous networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 135–144 (2017)
6. Lipton, Z.C., Kale, D.C., Elkan, C., Wetzel, R.: Learning to diagnose with lstm recurrent neural networks. arXiv preprint arXiv:1511.03677 (2015)
7. Liu, C., Wang, F., Hu, J., Xiong, H.: Temporal phenotyping from longitudinal electronic health records: A graph based framework. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 705–714 (2015)
8. Malinchoc, M., Kamath, P.S., Gordon, F.D., Peine, C.J., Rank, J., Ter Borg, P.C.: A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. Hepatology **31**(4), 864–871 (2000)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
10. Miotto, R., Li, L., Kidd, B.A., Dudley, J.T.: Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports **6**(1), 1–10 (2016)
11. Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. IEEE journal of biomedical and health informatics **22**(5), 1589–1604 (2017)
12. Zhu, Z., Yin, C., Qian, B., Cheng, Y., Wei, J., Wang, F.: Measuring patient similarities via a deep architecture with medical concept embedding. In: 2016 IEEE 16th International Conference on Data Mining (ICDM). pp. 749–758. IEEE (2016)