Hypergraph Transformers for EHR-based Clinical Predictions

Ran Xu[1], Mohammed K Ali, MD, MBA[2], Joyce C. Ho, PhD[1], Carl Yang, PhD[1]
[1] Department of Computer Science, Emory University, Atlanta, GA
[2] Hubert Department of Global Health, Emory University, Atlanta, GA

Abstract

Electronic health records (EHR) data contain rich information about patients' health conditions including diagnosis, procedures, medications and etc., which have been widely used to facilitate digital medicine. Despite its importance, it is often non-trivial to learn useful representations for patients' visits that support downstream clinical predictions, as each visit contains massive and diverse medical codes. As a result, the complex interactions among medical codes are often not captured, which leads to substandard predictions. To better model these complex relations, we leverage hypergraphs, which go beyond pairwise relations to jointly learn the representations for visits and medical codes. We also propose to use the self-attention mechanism to automatically identify the most relevant medical codes for each visit based on the downstream clinical predictions with better generalization power. Experiments on two EHR datasets show that our proposed method not only yields superior performance, but also provides reasonable insights towards the target tasks.

Introduction

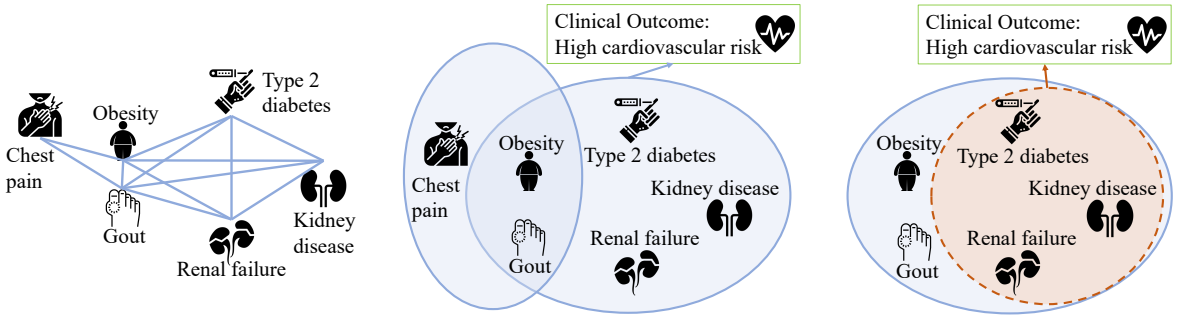
Nowadays, electronic health records (EHR) data are routinely collected across diverse healthcare institutions from millions of patients. EHR contain rich information about patients such as diagnosis, medication and lab results, and have been widely used to identify patterns for patients and assist with clinical decisions.[1] In recent years, there is a strong interest to apply machine learning techniques to support digital medicine.[2] At the individual level, patient's health records can foster personalized disease diagnosis and medication recommendation;[3] across populations, EHRs can provide a vital resource to understand population health management and help to design better healthcare policies.[4]

Despite its tremendous importance, one of the critical challenges for EHR modeling is to convert the patient's visit data to *meaningful* representations[1] for various kinds of medical codes and visits.[5] Traditional techniques often rely on expert-defined rules and feature engineering,[6,7] which can be labor intensive, and the resulting models often have limited generalizability across datasets or institutions. With the development of deep neural networks (DNN), several works directly learn distributed, low-dimensional embedding vectors for different medical codes.[8,9] However, these works ignore interactions among medical codes and yield vectors with limited representational power. Recently, graph neural networks (GNNs) have been proposed to learn over graph structures constructed from EHR data, in order to model the relations among patient visits and medical codes.[10] However, they only consider *pairwise* relations among medical codes, which is still insufficient to characterize the complex relationships between visits and medical codes.

In reality, a visit typically contains a large set of medical codes including diagnosis, medication, and procedure codes with varying sizes; and each medical code can also appear across a set of visits. Thus, modeling the interactions among medical codes using the vanilla graph with pairwise edges would lead to very dense graphs as illustrated in Figure 1a. Moreover, such graphs cannot capture which group of medical codes are present in the same visit. To the best of our knowledge, existing work has not considered appropriate data structures to preserve the complex relations among medical codes and visits, which is beneficial for learning embeddings for these units to support prediction tasks.

In this work, our central insight is to use *hypergraph*, a more flexible and general graph-based data structure, to model the high-order interactions and produce better medical code and patient visit representations for clinical predictions. Specifically, we view each visit as a hyperedge and each medical code as a node, and each hyperedge connects all the nodes that appear in the corresponding visit. Figure 1b shows an example of hypergraph construction on EHR data with two hospital visits. Compared with the vanilla graph construction where the degree for all edges is fixed to 2 as shown in Figure 1a, the hypergraph modeling has two main advantages: (1) *Preserving high-order interactions*: Hypergraph

[1] We use the term 'representation' and 'embedding' interchangeably in the remaining of the paper.



(a) Vanilla Graph Construction (b) Hypergraph Construction (c) Key Factors within the Visit

Figure 1: An example of encoding co-occurrence relationships in EHR with (a) vanilla graph and (b) hypergraph. Blue lines in (a) denote normal pairwise edges, blue circles in (b) denote hyperedges, and the red circle in (c) denotes the key factors within the hyperedge that are essential for the outcome prediction.

is capable of modeling high-order relationships beyond pairwise using its degree-free hyperedges. We argue that harnessing this property is critical for the downstream tasks, as combinatorial properties among medical codes can be relevant to a specific visit and its corresponding clinical outcomes. For example, Figure 1b shows that in a hospital visit the patient is diagnosed with five diseases - type2 diabetes, kidney disease, renal failure, obesity and gout. These five diseases all together indicate the patient's high risk of having cardiovascular disease in the future. Through the usage of hypergraphs, these different codes and visits can be collectively represented. (2) *Explicitly modeling visits and medical codes*: Hypergraphs preserve both medical code and visit information, while vanilla graphs only model interactions among medical codes. As a result, by using hypergraphs, we are able to preserve more information for each clinical visit, which can potentially benefit the visit-level prediction tasks.

With the hypergraph structure, we care about how to leverage its rich information for reasonable clinical predictions. Although there exist several neural architectures for hypergraphs,[11,12,13] even with an application on EHR data,[14] these methods mainly learn medical code representations by aggregating information from the hyperedge-connected neighbors, which are less useful for predictions on the visit level. Moreover, they often use the *mean pooling* method to aggregate visit and medical code embeddings, which treats different nodes in the hyperedges equally. In EHR data, each visit contains multiple medical codes, but not all of them are equally important for the clinical predictions.

Motivated by the challenges above, we propose HypEHR, a hypergraph-based neural network model, to accurately predict patients' clinical outcomes. Specifically, we leverage *set transformers*[15,11] to directly model both nodes and hyperedges as a function of its connected hyperedges and nodes, respectively. As set functions allow input sets of any size, we are able to flexibly encode the full high-order relations among nodes and hyperedges. Furthermore, we employ the self-attention mechanism in transformers to identify the most relevant medical codes for the visit. This not only helps the model capture more task-aware information and filter out spurious noises, but also provides valuable insights towards the prediction tasks. For example, based on the attention weights, we can extract type 2 diabetes, kidney disease and renal failure as key factors, since kidney disease, diabetes and heart disease are highly correlated [2] and renal failure is also known as an end-stage kidney disease.

We conduct experiments on two real-world EHR datasets, a publicly accessible de-identified dataset named MIMIC-III[16] and a privately-owned de-identified dataset from Emory Healthcare system named CRADLE, for phenotypes prediction and cardiovascular disease (CVD) risk prediction, respectively. The results illustrate that HypEHR achieves superior performance with the average gain of 2.59% in Area Under the Receiver Operating Characteristic curve (AUROC) and 4.00% in Area Under Precision/Recall curve (AUPR). Furthermore, the attention mechanism used in our set transformer has additional benefits: when evaluated by a domain expert, HypEHR generates higher weights for medical codes that are more relevant to the target tasks, justifying its efficacy on providing clinically useful insights.

[2] https://www.cdc.gov/kidneydisease/publications-resources/link-between-ckd-diabetes-heart-disease.html#:~:text=When%20the%20kidneys%20don't,can%20lead%20to%20heart%20disease

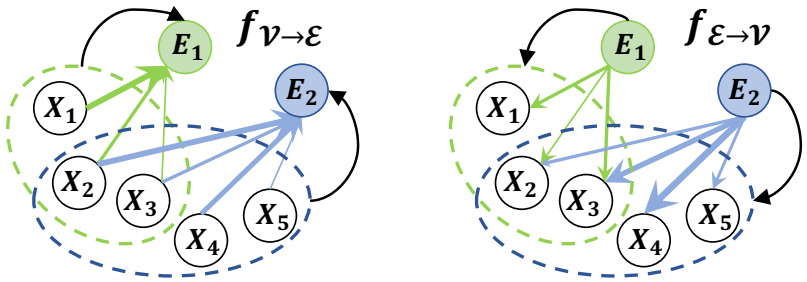


Figure 2: The framework of HypEHR. Two key steps include $f_{\mathcal{V}\to\mathcal{E}}$ which aggregates node information to learn hyperedge embeddings, and $f_{\mathcal{E}\to\mathcal{V}}$, which aggregates hyperedge embeddings to learn node embeddings. The *thickness* of the arrows indicate the learnable attention weights during the information aggregations.

Related Works

Learning to represent different medical codes in a unified space to support the downstream prediction tasks is an important topic for health informatics. With the development of DNNs, researchers started to explore the possibility of efficient representation learning for clinical elements in EHR. Earlier research borrowed the idea from text embedding models[17,18,19] to learn dense representations for medical concepts[20,21] to support clinical predictions. However, these approaches learn embeddings without explicitly modeling the relations among the different components.

More recently, graph-based models have been proposed for EHR modeling. They first build a co-occurrence graph from the EHR data, and then leverage graph neural networks (GNNs) to learn the relations among medical codes within each encounter for clinical outcome prediction.[10] However, their graph structures are usually predefined with domain expertise,[22,23] or prior knowledge,[24] which can be expensive to obtain and are less generalizable. Besides, the GNNs used in their studies are only able to encode pairwise relations, which is not ideal in EHR modeling, given the large set of medical codes involved in each visit. More related to us, hypergraph-based neural networks have been proposed to model the higher-order relationships among elements.[12,13,25] However, little attention has been paid to adopting these techniques for supporting medical prediction tasks. In the following section, we will introduce HypEHR, and illustrate how we leverage hypergraph neural networks to design accurate clinical predictive models.

Method

In this section, we propose HypEHR, our hypergraph-based framework for EHR modeling. We first introduce the definition of the problem. Then, we describe the framework to transform EHR data to hypergraphs for encoding the co-occurrence relationships among visits and medical codes. After that, we design a set transformer-based neural network to learn the relations among different medical codes and visits in the representation spaces and support downstream clinical prediction tasks. The overall process of our framework is shown in Figure 2.

Hypergraph Construction The EHR data used in this study comprises multiple types of medical codes $\mathcal{C}$, including *diseases*, *medications*, *procedures* and *services*. For each visit record, the *input* $\mathcal{X} \subset \mathcal{C}$ is a set of medical codes involved in the visit. We notice that EHR data have a characteristic that each visit contains massive medical codes for each visit and each medical code appears in multiple visits. While existing works often build a fully-connected graph[22,26] for modeling co-occurrence relationships, these works only encode pairwise relationships and cannot capture the complete hypergraph information. Furthermore, such fully-connected graphs are often too dense and can cause the memory-inefficient issue during the model training.

To better characterize the visit-level information, *hypergraph* structure is more beneficial in our setting as it can model their high-order interactions by jointly modeling hyperedges and nodes, and project both of them into an unified low-dimension space to facilitate prediction tasks. To transform the EHR into hypergraphs, we view each clinical visit as a hyperedge and each medical code as a node. Each hyperedge connects all the nodes appeared in the corresponding visit. We denote the hypergraph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}, \mathcal{E}$ stand for nodes and hyperedges, respectively.

Hypergraph Learning Since we aim to make a clinical prediction via the given visit, we seek to learn a better representation for hyperedges. While there exist various models for hypergraph representation learning,[25,12] these models mainly focus on *node-level* representation learning without dedicated designs for learning hyperedge embeddings. Thus, we design a neural network-based model to jointly learn *node* and *hyperedge* embeddings. In particular, our proposed hypergraph neural network comprises of L layers. In each layer, the update rule can be summarized as

$$\boldsymbol{E}_e^{(l)} = f_{\mathcal{V}\to\mathcal{E}}\left(\mathcal{V}_{e,\boldsymbol{X}^{(l-1)}}\right),\ \boldsymbol{X}_v^{(l)} = f_{\mathcal{E}\to\mathcal{V}}\left(\mathcal{E}_{v,\boldsymbol{E}^{(l-1)}}\right), \tag{1}$$

where $\boldsymbol{E}_e^{(l)}$ and $\boldsymbol{X}_v^{(l)}$ denote the embeddings of hyperedge e and node v in the l-th layer ($1 \le l \le L$), respectively. $\mathcal{V}_{e,\boldsymbol{X}}$ is the hidden representations of node that contain the hyperedge e, and $\mathcal{E}_{v,\boldsymbol{E}}$ is the hidden representations of hyperedges that contain the node v. In other words, hyperedge embeddings are obtained by aggregating information from nodes within each hyperedge, and similarly, node embeddings are derived by aggregating information from hyperedges that connect with the node. These two message passing processes are conducted iteratively as shown in Figure 2. In this way, we explicitly model the hypergraph embeddings to support the clinical prediction tasks.

To realize the two message passing function $f_{\mathcal{V}\to\mathcal{E}}(\cdot)$ and $f_{\mathcal{E}\to\mathcal{V}}(\cdot)$ (we use $f(\cdot)$ to represent these two functions in the following), several previous works only use mean pooling techniques to aggregate neighborhood information.[12,23,25] However, for hypergraphs induced from EHR data, we argue that there can exist a fraction of nodes for each hyperedge that is irrelevant to target prediction tasks. As a result, it often requires greater care to filter the nodes that are irrelevant to target prediction tasks to ensure the discriminative power of the hyperedge embeddings.

Towards this end, we leverage multi-head self-attention[27,28] to automatically identify the most important elements from neighbors during propagation.[29] Given the matrix $\boldsymbol{S} \in \mathbb{R}^{|\mathcal{S}|\times F}$ which represents the embedding of nodes or hyperedges $\mathcal{S}$ (e.g. $\mathcal{V}_e$ and $\mathcal{E}_v$ in Eq. 1) of F-dimensional vectors, the output of MultiHead attention $\boldsymbol{y} = f(\mathcal{S}) = \mathrm{MultiHead}(\boldsymbol{S}) \in \mathbb{R}^{1\times d}$ can be expressed as

$$\mathrm{MultiHead}(\boldsymbol{S}) = \|_{i=1}^{h}\mathbf{O}^{(i)} = \|_{i=1}^{h}\mathrm{SA}_i(\boldsymbol{S}), \quad \mathrm{SA}_i(\boldsymbol{S}) = \mathrm{softmax}\left(\frac{\mathbf{W}_i^Q(\boldsymbol{S}\mathbf{W}_i^K)^\top}{\sqrt{\lfloor d/h\rfloor}}\right)\boldsymbol{S}\mathbf{W}_i^V,$$

where $\boldsymbol{W}_i^Q \in \mathbb{R}^{1\times\lfloor d/h\rfloor}, \boldsymbol{W}_i^K \in \mathbb{R}^{F\times\lfloor d/h\rfloor}, \boldsymbol{W}_i^V \in \mathbb{R}^{F\times\lfloor d/h\rfloor}$ are learnable weight matrices, h is the number of attention heads, d is the dimension of the output vector. In this way, we automatically assign higher weights to the most relevant elements within the set of hyperedges or node with learnable parameters, thereby improving the discriminative power of the embeddings. By stacking L set transformer layers, we obtain the embeddings of nodes and hyperedges at the last layer as $\boldsymbol{X}^{(L)}$ and nodes as $\boldsymbol{E}^{(L)}$ respectively.

HypEHR incorporates additional techniques, namely *jumping knowledge* (JK)[30] and *PairNorm* (PN).[31] These techniques are introduced to combat the oversmoothing phenomenon.[32,33] Since EHR-based graphs tend to be dense, two hyperedges with several shared edges maybe indistinguishable from one another but in reality should be quite different. For example, take a patient with `type 2 diabetes` and `metformin` versus a patient with `type 2 diabetes`, `metformin`, and `high blood pressure`. The second will be at higher risk of cardiovascular disease and the embeddings should reflect this difference. Specifically, JK stacks embeddings from different transformer layers for downstream predictions, and PN adds a normalization layer for the learned node and hypergraph embeddings to prevent the embeddings from being identical.

Experimental Settings

Datasets We adopt a publicly accessible dataset named **MIMIC-III**[16] that comprises over forty thousand de-identified patients in critical care units of the Beth Israel Deaconess Medical Center from 2001 to 2012. We conduct **phenotyping prediction** on MIMIC-III and formulate it as a 25-class multi-label classification problem.[34] Specifically, it is to predict whether the 25 acute care conditions as described in the first two column of Table 3 will be present in patients' next visits, given their current ICU stay records. We identify these 25 phenotypes using Clinical Classifications Software (CCS) from the Healthcare Cost and Utilization Project (HCUP)[3]. Among them, there are 12 acute conditions, 8 chronic conditions and 5 mixed conditions, which are recurring acute diseases. During preprocessing,

[3]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixASingleDX.txt

we extract patients who have more than one hospital visit and get all pairs of adjacent visits for each patient. For each pair, we set the former one as input and the phenotypes in the latter one as labels. This gives us 12353 hyperedges with labels. Then we leverage the graph structure on these 12353 hyperedges with labels, as well as other 36875 hyperedges without labels from patients with only one encounter, to train the model.

Additionally, we use Project **CRADLE** (Emory Clinical Research Analytics Data Lake Environment), a privately-owned database that contains de-identified electronic health records at Emory Healthcare from 2013 to 2017. We focus on the patients with type 2 diabetes and predict whether those patients will experience **cardiovascular disease** (CVD) endpoints within a year after the initial diabetes diagnosis. The CVD endpoint is defined as the presence of coronary heart disease (CHD), congestive heart failure (CHF), myocardial infarction (MI), or Stroke, which are identified by their ICD-9 and ICD-10 clinical codes. We set the outcome label to be 1 if the patient have a CVD complication within a year, and 0 otherwise. Then for those positive patients, we preprocess the data to select the earliest recorded encounter within a year of the presence of the CVD endpoint as the input. And for those patients with label 0, we randomly select one encounter as the input from all the encounters that are at least one year before the last recorded encounter. Patients are excluded if they meet any of the following three criteria: (1) less than two encounters at Emory Healthcare, (2) the time interval from the first encounter to the last is less than one year, or (3) already have a history of a CVD condition. Eventually, there are 36611 patients remaining for the study. The sample characteristics of the two datasets are presented in existing works,[16,35] and detailed statistics are shown in Table 1. For both datasets, we use the ratio of 7:1:2 to split them into train/validation/test set.

Stats	**MIMIC-III**	**CRADLE**
# of diagnosis	846	7915
# of medication	4525	489
# of procedure	2032	4321
# of service	20	—
# of hyperedges	36875/12353	36611
Prevalence	see Table 3	21.4%

Table 1: Dataset statistics. For # of hyperedges in MIMIC-III, the first number indicates the hyperedges without labels, while the second one indicates ones with labels.

Evaluation Metrics. According to the label distribution shown in Table 1 and 3, both MIMIC-III and CRADLE are imbalanced. Thus we use accuracy, Area Under the Receiver Operating Characteristic curve (AUROC), Area Under Precision/Recall curve (AUPR) and macro-F1 score as the metrics.[14,22] For both accuracy and F1 score, we set 0.5 as the threshold to classify the predicted scores into different classes.

Implementation Details. We implement our model in PyTorch[4]. We use Adam as the optimizer with a learning rates of 1e-3. We set the weight decay to 1e-3, the number of set transformer layers $L = 3$, the hidden dimension $d = 48$, and the number of attention heads $h = 4$.

Baselines. We compare HypEHR with a comprehensive set of baselines[5]:

◇ **Non-graph Baselines.** These baselines learn the hidden relations within EHR data without leveraging graph structures. We select *Logistic Regression (LR)*,[36] *Support Vector Machine (SVM)*[37] and *Multi-layer Perceptron (MLP)*[38] in this category. For these methods, we first learn the embeddings for each medical code via Med2vec,[20] then pass the embedding with the model mentioned above for target classification tasks.

◇ **Graph-based Baselines.** These methods encode relations among different items in EHR using graphs. Specifically, they create an edge between two medical codes if they co-occur in a visit. We consider two baselines: *GCT*,[22] which proposes Graph Convolutional Transformer to learn the hidden EHR structure for predictive tasks, and *GAT*,[39] which uses attention-based massage passing mechanism for aggregating neighbor features. For these two methods, a task-specific MLP is stacked on the top of the model for prediction.

[4]https://pytorch.org/

[5]For other graph-based baselines, we also use JK and PN as additional techniques to ensure the fair comparison.

◇ **Hypergraph-based Baselines.** These baselines uses the same hypergraph structure as HypEHR but with different neural architectures for hypergraph learning. Specifically, we select two representative methods *HyperGCN*[12] and *HCHA*,[13] in our experiments. *HyperGCN* replaces each hyperedge of the hypergraph by a set of weighted pairwise edges connecting the vertices of the hyperedge to transform the hypergraph learning task to a vanilla graph learning problem. *HCHA* adopts clique based expansion to transform the hypergraph to a vanilla graph, and use attention weights conditioning on node features for message passing.

Experimental Results

Model	MIMIC-III				CRADLE			
	ACC	AUROC	AUPR	F1	ACC	AUROC	AUPR	F1
LR[36]	68.66 ± 0.24	64.62 ± 0.25	45.63 ± 0.32	13.74 ± 0.40	76.22 ± 0.30	57.22 ± 0.28	25.99 ± 0.26	42.18 ± 0.35
SVM[37]	72.02 ± 0.12	55.10 ± 0.14	34.19 ± 0.17	32.35 ± 0.21	68.57 ± 0.13	53.57 ± 0.11	23.50 ± 0.15	52.34 ± 0.22
MLP[38]	70.73 ± 0.24	71.20 ± 0.22	52.14 ± 0.23	16.39 ± 0.30	77.02 ± 0.17	63.89 ± 0.18	33.28 ± 0.23	45.16 ± 0.26
GCT[22]	76.58 ± 0.23	78.62 ± 0.21	63.99 ± 0.27	35.48 ± 0.34	77.26 ± 0.22	67.08 ± 0.19	35.90 ± 0.20	56.66 ± 0.25
GAT[39]	76.75 ± 0.26	78.89 ± 0.12	66.22 ± 0.29	34.88 ± 0.33	77.82 ± 0.20	66.55 ± 0.27	36.06 ± 0.18	56.43 ± 0.26
HyperGCN[12]	78.01 ± 0.23	80.34 ± 0.15	67.68 ± 0.16	39.29 ± 0.20	78.18 ± 0.11	67.83 ± 0.18	38.28 ± 0.19	60.24 ± 0.21
HCHA[13]	78.07 ± 0.28	80.42 ± 0.17	68.56 ± 0.15	37.78 ± 0.22	78.60 ± 0.15	68.05 ± 0.17	39.23 ± 0.13	59.26 ± 0.21
HypEHR	**79.07 ± 0.31***	**82.19 ± 0.13***	**71.08 ± 0.17***	**41.51 ± 0.25***	**79.76 ± 0.18***	**70.07 ± 0.13***	**40.92 ± 0.12***	**61.23 ± 0.18***

Table 2: Performance on MIMIC-III and CRADLE compared with different baselines. The result is averaged over 5 runs. We use * to indicate statistical significant results ($p < 0.05$).

Phenotype	Type	Prevalence	HypEHR		GAT		MLP	
			AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
Acute and unspecifed renal failure	acute	29.3%	67.19	45.31	63.24	40.63	56.35	35.42
Acute cerebrovascular disease	acute	7.1%	58.11	11.16	53.4	8.69	49.06	7.22
Acute myocardial infarction	acute	7.4%	73.56	20.52	66.31	14.99	51.69	8.77
Cardiac dysrhythmias	mixed	43.2%	76.61	72.57	72.79	68.91	55.88	49.89
Chronic kidney disease	chronic	25.5%	86.42	74.33	78.37	60.52	61.11	32.99
Chronic obstructive pulmonary disease	chronic	18.0%	82.16	56.15	78.24	46.88	55.57	20.52
Complications of surgical/medical care	acute	84.0%	70.65	91.96	66.63	90.50	64.87	89.79
Conduction disorders	mixed	2.4%	63.83	4.08	61.39	3.99	57.25	3.86
Congestive heart failure; nonhypertensive	mixed	39.2%	82.52	74.95	79.27	69.59	54.62	44.16
Coronary atherosclerosis and related	chronic	29.6%	82.31	69.56	77.15	60.71	56.81	39.00
Diabetes mellitus with complications	mixed	36.2%	92.80	88.63	89.72	85.12	56.16	40.14
Diabetes mellitus without complication	chronic	42.1%	86.80	85.48	83.89	79.67	56.75	46.57
Disorders of lipid metabolism	chronic	27.6%	76.65	54.52	72.64	47.29	55.24	34.31
Essential hypertension	chronic	35.4%	76.15	64.26	71.75	57.97	51.22	38.58
Fluid and electrolyte disorders	acute	44.4%	64.12	59.25	63.35	57.86	60.39	53.22
Gastrointestinal hemorrhage	acute	27.8%	65.18	43.51	62.44	39.99	53.42	29.08
Hypertension with complications	chronic	59.5%	76.64	80.99	74.32	77.83	56.28	63.85
Other liver diseases	mixed	21.9%	68.75	45.90	64.01	40.46	56.55	24.19
Other lower respiratory disease	acute	35.4%	67.16	55.57	64.93	52.95	57.18	42.50
Other upper respiratory disease	acute	9.5%	63.81	26.87	58.64	22.24	54.88	11.22
Pleurisy; pneumothorax; pulmonary collapse	acute	33.9%	65.50	49.20	64.12	47.89	57.10	42.05
Pneumonia	acute	21.7%	63.85	32.05	59.72	26.21	57.38	30.21
Respiratory failure; insufciency; arrest	acute	32.8%	66.31	51.30	64.38	48.66	56.76	41.09
Septicemia (except in labor)	acute	26.5%	64.26	37.85	60.91	35.68	60.36	35.38
Shock	acute	12.2%	62.74	17.93	59.83	17.79	60.67	18.04

Table 3: Description and label distribution of the 25 phenotypes in MIMIC-III and their associated model performance.

Table 2 summarizes the experimental results on the two datasets. Note that since accuracy and F1 are influenced by the threshold used for separating predicted scores into different classes, they are *less comprehensive* than AUROC and AUPR in demonstrating model performance. From the results, we have the following findings:

Visit Information	(1) Disorder of skin and/or subcutaneous tissue (2) Pure hypercholesterolemia (3) Office or other outpatient visit for the evaluation and management of an established patient (4) Degeneration of intervertebral disc (5) Essential hypertension (6) Primary malignant neoplasm of prostate (7) Aspirin received within 24 hours before emergency department arrival or during emergency department stay (EM) (8) Computed tomography, head or brain; without contrast material (9) Low back pain (10) Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only (11) Frontotemporal dementia (12) Opioids in combination with non-opioid analgesics (13) Cholelithiasis without obstruction (14) Tremor (15) Other opioids (16) Emergency department visit for the evaluation and management of a patient. (17) Nausea (18) Dizziness and giddiness (19) Type 2 diabetes mellitus without complication (20) Chest x-ray
Label	1 (The patient would experience CVD complications in the next year.)

Figure 3: The first case study: the green highlighted items are the ones with high attention weights.

◇ HypEHR outperforms all the baselines over four different evaluation metrics on both datasets. Compared to the best baselines, HypEHR raises the performance by 2.59% in AUROC and 4.00% in AUPR. This verifies that jointly learning both node and hyperedge embeddings is better than only tuning node features towards clinical outcome prediction tasks, as visit information is more related to the outcome labels. Moreover, leveraging self-attention mechanism during neighborhood aggregation enhances the model performance compared to the mean pooling used in previous works (HyperGCN), as it is able to automatically identifying critical medical codes in each visit for the target task.

◇ Graph-based models generally have a better performance than traditional machine learning methods. This phenomenon demonstrates that explicitly considering the interaction between nodes via message passing is beneficial for EHR modeling. We also observe that GAT has a slightly better performance than GCT, which verifies that the attention mechanism that focuses on the most relevant parts of the input raises the performance. However, the general performance gain of graph-based methods over traditional machine learning methods is limited, indicating that there exists better ways for encoding co-occurrence relationships.

◇ Hypergraph-based models can further improve the prediction performance over the graph-based models. This illustrates that modeling higher-order interactions beyond pairwise contributes to performance gain, especially on clinical EHR data, as it naturally involves a great number of medical codes in each visit. We notice that HCHA also leverages attention mechanism, but it cannot outperform HyperGCN. This is because their attention weights are calculated based on the node and hyperedge attributes, but the hyperedge attributes are not provided in our EHR data. Thus, the attention module in HCHA is less powerful when applying to EHR datasets.

Moreover, we summarize the per-task performance results for the 25 phenotypes in MIMIC-III in Table 3. For illustration purposes, we only present MLP and GAT but the performance trends of the other models follow the results in Table 2. We observe that HypEHR outperforms the other two baseline models on all the 25 phenotypes in terms of AUROC and AUPR, and GAT outperforms MLP on 23 phenotypes out of the total 25. The phenomenon demonstrates that our gain is consistent over almost all the phenotyping tasks, and again illustrates that graph-based methods have more expressive power than traditional machine learning methods.

Case Studies

To illustrate that self-attention mechanism is able to provide some valuable insights towards the clinical outcome, we analyze the important medical codes (i.e., high attention weights) in each visit to demonstrate that they identify key factors that are relevant to the clinical outcome. First, we randomly sample a handful of visits from the CRADLE

Visit Information	(1) Office or other outpatient visit for the evaluation and management of an established patient (2) Essential hypertension (3) Obesity (4) Hyperlipidemia (5) Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report (6) Dipeptidyl peptidase 4 (DPP-4) inhibitors (7) Hypertensive complication (8) Cardiovascular stress test (9) Combinations of various lipid modifying agents (10) Preoperative cardiovascular examination (11) Antiinflammatory agents, non-steroids (12) ACE inhibitors and calcium channel blockers (13) Antiinflammatory preparations, non-steroids for topical use (14) Tricuspid valve disorder, non-rheumatic (15) Insulins and analogues for injection, intermediate- or long-acting combined with fast-acting (16) Office consultation for a new or established patient (17) Type 2 diabetes mellitus without complication (18) Pre-surgery evaluation (19) Echocardiography, transthoracic, real-time with image documentation (2D) (20) Acetic acid derivatives and related substances (21) Other cardiac preparations (22) Myocardial perfusion imaging, tomographic (SPECT) (23) Biguanides (24) Preparations inhibiting uric acid production (25) Combinations of oral blood glucose lowering drugs
Label	1 (The patient would experience CVD complications in the next year.)

Figure 4: The second case study: the green highlighted items are the ones with high attention weights.

dataset. We then sort the attention weight into descending order and select the top 30% nodes as the extracted subset. With the visit information and the extracted elements, a *medical domain expert* evaluated by the key factors and the CVD outcome. We present two cases (see Figure 3 and 4) to demonstrate the quality of HypEHR's explanations.

According to the analysis provided by the domain expert, the subsets in both Figure 3 and 4 extracted by attention weight from the original set contain many key factors that indicate the patient's CVD condition in the next year. Specifically, in the first case (Figure 3), the extracted record shows that the patient had frontotemporal dementia, received aspirin before arriving at the emergency department and had a CT of the head. Frontotemporal demetia (FTD), as shown by existing studies,[40,41] has a specific association with cardiovascular disease. Moreover, the CT of the head was performed to better understand the patients' FTD severity. Additionally, aspirin indicates a high risk of atherosclerosis.[42] Thus, these factors imply that the patient is likely to have Stroke or myocardial infarction (MI), which are complications captured in the CVD endpoint.

In the second case, the patient is on an insulin regimen. Patients on this regimen are predisposed to develop many CVD related diseases such as inflammation, atherosclerosis, hypertension, dyslipidemia, heart failure (HF), and arrhythmias.[43] The stress testing, echocardiography and myocardial perfusion imaging extracted by the attention weights also suggest that the patient might already experience some symptoms associated with CVD,[44,45] but has not been diagnosed formally with CVD.

To summarize, the medical codes with higher attention weights can be considered as key factors and provide meaning insights towards the clinical prediction tasks. The rationality of these extracted subsets are verified by a domain expert, and supported by existing medical literature, and thus can be potentially applied to assist medical decision making.

Conclusion

In this work, we propose HypEHR, a hypergraph-based neural network model to accurately predict patients' clinical outcomes. Specifically, we leverage hypergraph to effectively model the complex interactions among patients' visits and medical conditions. Such a hypergraph structure goes beyond the pairwise relationships used in existing graph-based EHR models and can flexibly encode medical conditions at the visit level. Furthermore, we employ set

transformers with self-attention mechanism to automatically identify the key factors within each visit to learn accurate hyperedge and node representations and provide insight into the clinical predictions. The experiments on two real-world EHR datasets demonstrate that HypEHR outperforms the best existing baseline by 2.59% in AUROC and 4.00% in AUPR. In addition, as has been verified by a domain expert, the attention weights can be used to provide valuable insights towards the target clinical predictions. For future works, we consider including chronological information or unstructured data for better generalization. We also consider modeling various node types as a heterogeneous graph.

Acknowledgements

This research was partially supported by the internal funds and GPU servers provided by the Computer Science Department of Emory University. MKA was partially supported by the Georgia Center for Diabetes Translation Research, funded by the National Institute of Diabetes Digestive and Kidney Disorders (P30DK111024). JCH was supported by NSF grants IIS-1838200, IIS-2145411 and NIH grant 5K01LM012924-03.

References

1. King J, Patel V, Jamoom EW, Furukawa MF. Clinical benefits of electronic health record use: national findings. Health services research. 2014;49(1pt2):392-404.
2. Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. NPJ digital medicine. 2018;1(1):1-4.
3. Dong X, Rashidian S, Wang Y, Hajagos J, Zhao X, Rosenthal RN, et al. Machine learning based opioid overdose prediction using electronic health records. In: AMIA Annual Symposium Proceedings; 2019. p. 389.
4. Landi I, Glicksberg BS, Lee HC, Cherng S, Landi G, Danieletto M, et al. Deep representation learning of electronic health records to unlock patient stratification at scale. NPJ digital medicine. 2020;3(1):1-11.
5. Si Y, Du J, Li Z, Jiang X, Miller T, Wang F, et al. Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. Journal of Biomedical Informatics. 2021;115:103671.
6. Chen R, Georgii-Hemming P, Åhlfeldt H. Representing a chemotherapy guideline using openEHR and rules. In: Medical Informatics in a United and Healthy Europe. IOS Press; 2009. p. 653-7.
7. Yu Y, Zuo S, Jiang H, Ren W, Zhao T, Zhang C. Fine-Tuning Pre-trained Language Model with Weak Supervision: A Contrastive-Regularized Self-Training Approach. In: NAACL-HLT; 2021. p. 1063-77.
8. Agarwal K, Eftimov T, Addanki R, Choudhury S, Tamang S, Rallo R. Snomed2Vec: Random Walk and Poincar\'e Embeddings of a Clinical Knowledge Base for Healthcare Analytics. arXiv preprint arXiv:190708650. 2019.
9. Choi Y, Chiu CYI, Sontag D. Learning low-dimensional representations of medical concepts. AMIA Summits on Translational Science Proceedings. 2016;2016:41.
10. Ochoa JGD, Mustafa FE. Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses. Artificial Intelligence in Medicine. 2022:102359.
11. Chien E, Pan C, Peng J, Milenkovic O. You are AllSet: A Multiset Function Framework for Hypergraph Neural Networks. In: International Conference on Learning Representations; 2022. .
12. Yadati N, Nimishakavi M, Yadav P, Nitin V, Louis A, Talukdar P. Hypergcn: A new method for training graph convolutional networks on hypergraphs. Advances in neural information processing systems. 2019;32.
13. Bai S, Zhang F, Torr PH. Hypergraph convolution and hypergraph attention. Pattern Recognition. 2021;110:107637.
14. Cai D, Sun C, Song M, Zhang B, Hong S, Li H. Hypergraph Contrastive Learning for Electronic Health Records. In: Proceedings of the 2022 SIAM International Conference on Data Mining (SDM). SIAM; 2022. p. 127-35.
15. Lee J, Lee Y, Kim J, Kosiorek A, Choi S, Teh YW. Set transformer: A framework for attention-based permutation-invariant neural networks. In: International conference on machine learning. PMLR; 2019. p. 3744-53.
16. Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3(1):1-9.
17. Yu Y, Xiong C, Sun S, Zhang C, Overwijk A. COCO-DR: Combating Distribution Shifts in Zero-Shot Dense Retrieval with Contrastive and Distributionally Robust Learning. arXiv preprint arXiv:221015212. 2022.
18. Meng Y, Huang J, Wang G, Zhang C, Zhuang H, Kaplan L, et al. Spherical text embedding. Advances in neural information processing systems. 2019.
19. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their

compositionality. Advances in neural information processing systems. 2013;26.
20. Choi E, Bahadori MT, Searles E, Coffey C, Thompson M, Bost J, et al. Multi-layer representation learning for medical concepts. In: proceedings of the 22nd ACM SIGKDD international conference; 2016. p. 1495-504.
21. Finch A, Crowell A, Bhatia M, Parameshwarappa P, Chang YC, Martinez J, et al. Exploiting hierarchy in medical concept embedding. JAMIA open. 2021;4(1):ooab022.
22. Choi E, Xu Z, Li Y, Dusenberry M, Flores G, Xue E, et al. Learning the graphical structure of electronic health records with graph convolutional transformer. In: Proceedings of AAAI. vol. 34; 2020. .
23. Yu Y, Xia T, Wang H, Feng J, Li Y. Semantic-aware spatio-temporal app usage representation via graph convolutional network. IMWUT. 2020;4(3).
24. Liu Z, Li X, Peng H, He L, Philip SY. Heterogeneous similarity graph neural network on electronic health records. In: Big Data. IEEE; 2020. p. 1196-205.
25. Feng Y, You H, Zhang Z, Ji R, Gao Y. Hypergraph neural networks. In: Proceedings of AAAI; 2019. p. 3558-65.
26. Zhu W, Razavian N. Variationally regularized graph-based representation learning for electronic health records. In: Proceedings of the Conference on Health, Inference, and Learning; 2021. p. 1-13.
27. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.
28. Yu Y, Huang K, Zhang C, Glass LM, Sun J, Xiao C. SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization. Bioinformatics. 2021;37(18):2988-95.
29. Xu R, Yu Y, Cui H, Kan X, Zhu Y, Ho JC, et al. Neighborhood-regularized Self-training for Learning with Few Labels. In: Proceedings of AAAI; 2023. .
30. Xu K, Li C, Tian Y, Sonobe T, Kawarabayashi Ki, Jegelka S. Representation learning on graphs with jumping knowledge networks. In: International conference on machine learning. PMLR; 2018. p. 5453-62.
31. Zhao L, Akoglu L. PairNorm: Tackling Oversmoothing in GNNs. In: ICLR; 2020. .
32. Chen D, Lin Y, Li W, Li P, Zhou J, Sun X. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In: Proceedings of the AAAI Conference; 2020. p. 3438-45.
33. Yu, Li Y, Shen J, Feng H, Sun J, Zhang C. Steam: Self-supervised taxonomy expansion with mini-paths. In: the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2020. p. 1026-35.
34. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Scientific data. 2019;6(1):1-18.
35. Ho JC, Staimez LR, Narayan KV, Ohno-Machado L, Simpson RL, Hertzberg VS. Evaluation of available risk scores to predict multiple cardiovascular complications for patients with type 2 diabetes mellitus using electronic health records. Computer Methods and Programs in Biomedicine Update. 2023;3:100087.
36. Menard S. Applied logistic regression analysis. 106. Sage; 2002.
37. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273-97.
38. Naraei P, Abhari A, Sadeghian A. Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data. In: 2016 Future Technologies Conference (FTC). IEEE; 2016. .
39. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks. In: International Conference on Learning Representations; 2018. .
40. Irimata KE, Dugger BN, Wilson JR. Impact of the presence of select cardiovascular risk factors on cognitive changes among dementia subtypes. Current Alzheimer Research. 2018;15(11):1032-44.
41. Golimstok A, Cámpora N, Rojas JI, Fernandez MC, Elizondo C, Soriano E, et al. Cardiovascular risk factors and frontotemporal dementia: a case–control study. Translational neurodegeneration. 2014;3(1):1-6.
42. Ittaman SV, VanWormer JJ, Rezkalla SH. The role of aspirin in the prevention of cardiovascular disease. Clinical medicine & research. 2014;12(3-4):147-54.
43. Herman ME, O'Keefe JH, Bell DS, Schwartz SS. Insulin therapy increases cardiovascular risk in type 2 diabetes. Progress in cardiovascular diseases. 2017;60(3):422-34.
44. Capotosto L, Massoni F, De Sio S, Ricci S, Vitarelli A. Early diagnosis of cardiovascular diseases in workers: role of standard and advanced echocardiography. BioMed Research International. 2018;2018.
45. Rana JS, Rozanski A, Berman DS. Combination of myocardial perfusion imaging and coronary artery calcium scanning: potential synergies for improving risk assessment in subjects with suspected coronary artery disease. Current atherosclerosis reports. 2011;13(5):381-9.