HeteroMed: Heterogeneous Information Network for Medical Diagnosis

Anahita Hosseini
University of California Los Angeles
Los Angeles, California
anahosseini@ucla.edu

Ting Chen
University of California Los Angeles
Los Angeles, California
tingchen@cs.ucla.edu

Wenjun Wu
University of California Los Angeles
Los Angeles, California
wenjunwu@ucla.edu

Yizhou Sun
University of California Los Angeles
Los Angeles, California
yzsun@cs.ucla.edu

Majid Sarrafzadeh
University of California Los Angeles
Los Angeles, California
majid@cs.ucla.edu

ABSTRACT

With the recent availability of Electronic Health Records (EHR) and great opportunities they offer for advancing medical informatics, there has been growing interest in mining EHR for improving quality of care. Disease diagnosis due to its sensitive nature, huge costs of error, and complexity has become an increasingly important focus of research in past years. Existing studies model EHR by capturing co-occurrence of clinical events to learn their latent embeddings. However, relations among clinical events carry various semantics and contribute differently to disease diagnosis which gives precedence to a more advanced modeling of heterogeneous data types and relations in EHR data than existing solutions.

To address these issues, we represent how high-dimensional EHR data and its rich relationships can be suitably translated into HeteroMed, a heterogeneous information network for robust medical diagnosis. Our modeling approach allows for straightforward handling of missing values and heterogeneity of data. HeteroMed exploits metapaths to capture higher level and semantically important relations contributing to disease diagnosis. Furthermore, it employs a joint embedding framework to tailor clinical event representations to the disease diagnosis goal. To the best of our knowledge, this is the first study to use Heterogeneous Information Network for modeling clinical data and disease diagnosis. Experimental results of our study show superior performance of HeteroMed compared to prior methods in prediction of exact diagnosis codes and general disease cohorts. Moreover, HeteroMed outperforms baseline models in capturing similarities of clinical events which are examined qualitatively through case studies.

CCS CONCEPTS

• **Computing methodologies** → **Ranking**; **Learning latent representations**; *Learning to rank*; *Unsupervised learning*; • **Applied computing** → **Health informatics**; *Health care information systems*; • **General and reference** → *Reference works*; • **Information systems** → *Learning to rank*;

KEYWORDS

Heterogeneous Information Network; Electronic Health Record; Health informatics; Network Embedding

ACM Reference Format:
Anahita Hosseini, Ting Chen, Wenjun Wu, Yizhou Sun, and Majid Sarrafzadeh. 2018. HeteroMed: Heterogeneous Information Network for Medical Diagnosis. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3269206.3271805

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3271805

1 INTRODUCTION

Electronic Health Records (EHR) provide detailed documented information on various clinical events that occur during a patient's stay in the hospital. Laboratory tests, medications, nurse notes, and diagnoses are examples of heterogeneous types of clinical records. Availability of EHR data in recent years has opened great opportunities for researchers to further explore computer-aided advancements in the healthcare domain. One goal of many existing studies is improving clinical decision making and disease diagnosis.

The diagnostic process involves careful consideration of clinical observations (such as symptoms and diagnostic tests), extraction of relevant information, and more importantly paying attention to their relations. A clinical observation is generally non-specific to a single disease. It is its relation or co-occurrence with other observations that can be indicative of a disease [16]. Moreover, the presence of multiple diseases can cause complexity in observations and their relations. These complexities along with a large amount of information to be analyzed by clinicians make their decisions prone to cognitive error and in many cases suboptimal, which can be very costly and in some cases fatal. A study on 100 diagnostic errors showed that cognitive factor contributed to 74% of the errors made [16]. Therefore, building a computer-aided diagnosis system is of great importance in reducing error and improving healthcare.

To design such system, obtaining a structured and informative model of the EHR data is necessary. Prior studies employ different approaches for this aim. A group of them employed feature engineering to represent clinical events in EHR and used deep or shallow models for the prediction goals [5, 14, 28]. However, high

dimensionality of EHR data, commonness of missing values, and need for extensive clinical knowledge are main challenges that arise in this approach and introduce many limitations. Others employed unsupervised representation learning of clinical events, patients, and visits [7, 10–12]. These methods that are mostly inspired by Med2vec [7], consider co-occurrence of clinical events in different patient records to extract latent embeddings of these entities. However, representations learned are general and not tailored to the goal of diagnosis prediction. More importantly, none of the abovementioned approaches can capture the rich structure of EHR data and semantics of various relations it contains. This information make a great contribution to the prediction goal and the model should be able to capture and reflect those into the learned representations. Therefore, any adopted EHR modeling approach should achieve two main goals:

- properly capture structure of EHR data and semantic of relations; and
- learn representations suitable for disease diagnosis goal.

To address these requirements and shortcomings of prior models, we propose a disease diagnosis model based on Heterogeneous Information Network (HIN) [18]. HINs, information networks with various types of nodes and relations, have gained lots of attention in recent years due to their ability in distinguishing and learning the different semantics of relations among entities [31] and can be profoundly beneficial to better express the rich network of patients and clinical events in the EHR data.

We introduce how EHR can be translated into an HIN and introduce our node extraction strategies from different formats of data (e.g., raw text, numerical, categorical) present in EHR. We then exploit metapaths [18] to introduce composite relation semantics into our network and capture those that are informative for our diagnostic purposes. We afterward employ a heterogeneous embedding model [13] to learn representations of clinical events of the network, which samples similar nodes from pre-defined metapaths. This allows our model to learn similarity of clinical events and patients with respect to semantically important paths rather than random sampling strategy used in prior skip-gram based diagnosis studies. To further tailor latent embeddings to diagnosis prediction goal, a supervised embedding model is jointly learned to adjust representations, using the framework proposed by [6]. While our diagnosis prediction model only utilizes diagnostic information for reasoning the disease, we propose exploiting the treatment information at the time of unsupervised representation learning to improve learned embeddings and capture similarity of clinical events in terms of outcome. Apart from relation-aware modeling and tailored representation learning, HINs also offer the advantage of straightforward handling of missing values, which is a common challenge in clinical data modeling.

We demonstrate that employing HIN for modeling EHR and diagnosis prediction outperforms state of the art models in two levels of general disease cohort and specific diagnosis prediction. We also conduct two case studies to qualitatively reveal the strength of HeteroMed in capturing relations in clinical data which are validated by a clinician. In short, contributions of this study are:

- We propose HeteroMed, an HIN-based medical model for disease diagnosis which captures the semantics of clinical entities and relations and learns embeddings tailored for disease diagnosis task.
- We demonstrate how EHR data can be translated into an HIN to achieve optimal learning power.
- We empirically show HeteroMed outperforms existing diagnostic models qualitatively and quantitatively.

2 RELATED WORK

To tackle the problem of disease prediction, initial studies until recent years employed conventional feature engineering methods to extract clinical representations and predict diseases [21, 33, 35]. However, feature engineering for clinical domain is a tedious task and requires expert knowledge. Moreover, missing values in EHR pose a great challenge to feature extraction [3]. A number of recent studies [5, 28] employ feature engineering approach along with a deep model to predict high level disease categories, which achieve improvements in results but still experience same issues.

Recognizing discussed challenges, recent studies employ unsupervised representation learning approaches [10, 14, 15]. Most of proposed models, inspired by success of word2vec [24, 25] in natural language processing, turn clinical event records into an ordered sequence of words and employ skip-gram [25] to capture co-occurrence of clinical events and learn latent embeddings [11, 12]. To expand the idea, Med2vec [7] proposes a multilayer representation learning model for clinical code and visit embedding which also initializes the embeddings using skip-gram but modifies them through the network training process. Although these models are successful in eliminating the need for clinical expert knowledge, they fail to capture EHR structure and its internal relations and can only learn a general-purpose representation. It is worth mentioning that the recent studies Dipole [23], RETAIN [9], and GRAM [8] although similar in subject, are different than ours. These studies are mainly focused on employing history of admissions and diagnoses for future disease prediction. While our study is focused on the diagnosis based on clinical events happening during a single admission.

Heterogeneous Information Networks [18] are different from homogeneous ones in their ability of representing multiple types of nodes and relations. This capability has attracted lots of attention in different applications such as personalized recommendation [32] and malware detection [19]. Due to the large size of real-world networks and sparsity of data, representing nodes as a low-dimensional vectors is a widely adopted approach in network mining. Network representation learning techniques in general are inspired by word2vec [24], among which DeepWalk [27], LINE [34], and Node2vec [17] have been utilized in many network mining researches. Recent studies have adopted similar techniques for heterogeneous network representation learning [4, 6, 13]. They include the heterogeneity of nodes in the definition of relations and neighbors. Furthermore, it is demonstrated in [6] that a joint embedding approach in heterogeneous node representation learning can lead to improved supervised task performance. In this study, we employ heterogeneous network embedding alongside with the joint learning framework to learn clinical event representations.

3 METHODOLOGY

In this section, we first put forward the problem definition and terminology used in the study. Then, we introduce how EHR can be viewed as a heterogeneous network and discuss network construction techniques. Lastly, we discuss the training and prediction models adopted for our disease diagnosis task.

3.1 Problem Definition and Clinical Terminology

Each record in EHR data is conventionally called a clinical event. A clinical event e can be viewed as a triple: $e = (t, n, v)$ where t, n, and v respectively denote *type*, *name*, and *value* of it. *Glucose level of 60* is an example of a clinical event that has type = laboratory test, name = Glucose, and value = 60. Clinical events based on their type may or may not have a value.

Furthermore, the set of clinical event types in EHR are denoted as $t_1, t_2, \ldots, t_{|T|} \in T$ where $T = DIAGNOSTIC \cup TREATMENT$ and $DIAGNOSTIC$ = {*laboratory test*, *symptom*, *age*, *gender*, *ethnicity*, *microbiology test*} and $TREATMENT$ = {*prescription*, *procedure*, *diagnosis*}. Diagnostic clinical events are the source of information for disease diagnosis. This is while clinical events in treatment category happen after the diagnostic process and should not be directly used for diagnosis prediction.

Therefore, having clinical events for a patient p represented as $E(p) = \{E_1(p), \ldots, E_T(p)\}$ where $E_t(p)$ denotes all type t clinical events recorded for p, we define the problem of disease diagnosis as prediction of p's diagnosis clinical events (D_p), denoted as $D_p = [E_t(p) \mid t = diagnosis]$, given the diagnostic clinical events of p: $\{E_t(p) \mid t \in DIAGNOSTIC\}$. Due to the large size of all possible diagnoses in EHR data, we define diagnosis prediction as a ranking problem such that top results of prediction model should ideally match the real diagnosis set (D_p) for patient p.

3.2 EHR from a Heterogeneous Network Point-of-View

Multiple types of clinical events and their various types of relations can be intuitively viewed as a heterogeneous network.

Definition 3.1. Heterogeneous Information Network is defined as a graph $G = (V, E)$ in which nodes and links between them can have various types. Nodes are mapped to their type by a node mapping function $g_v : V \rightarrow A$ where A is the set of all node types and similarly a link mapping function $g_e : E \rightarrow R$ maps links to their type where R is the set of all possible link types. By definition we have $|R| > 1$ or $|A| > 1$. Furthermore, $S_G = (A, R)$ denotes the network schema.

Different patients and clinical events form the nodes of our clinical heterogeneous network. The type of a node in this network is defined by the type of the clinical event mapped to it. Moreover, links of the network are designed based on the basic EHR relations which are mainly between a patient and a clinical event (e.g., patient's relation to his laboratory tests or symptoms). Figure 1 shows the abstract schema of the network illustrating node types and basic links. The figure also specifies if nodes belong to the treatment or diagnostic type category.

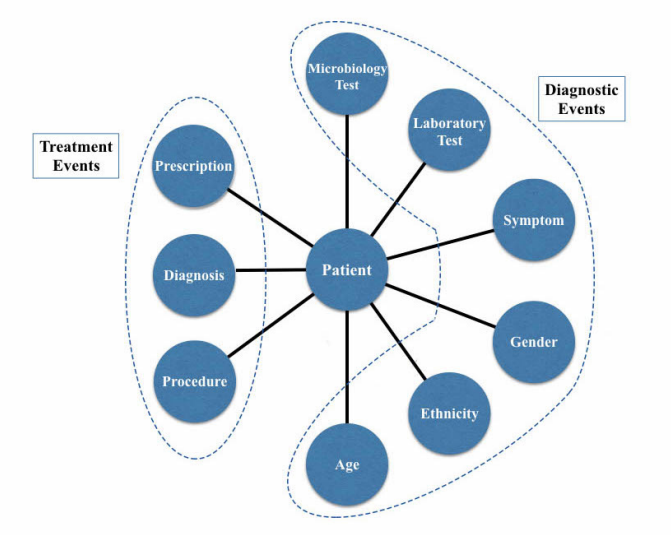

Figure 1: EHR heterogeneous network schema.

To further enrich the network with semantics of relation in EHR, new compositional relations can be defined using Metapaths [4].

Definition 3.2. Metapaths in HIN define higher order relations between two node types. Having the network schema $S_G = (A, R)$, a metapath schema is denoted as $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_m} A_{m+1}$.

A metapath is considered as a new link in the network and is added by creating a new connectivity between start and end nodes of any path matching the metapath schema. Metapaths allow our network to better learn the semantics of similarity among nodes. For instance, *patient*$\rightarrow$ *symptom* $\leftarrow$ *patient* captures similarity of patients in terms of their symptoms.

3.3 Construction of HIN from Clinical Events

In this section, we introduce the proper modeling approach for construction of HIN from EHR data and our technique in extraction of some clinical events from raw text.

In general, having a clinical event $e = (t, n, v)$, it can be mapped into a node of type t with identification of (n, v). For instance, a Glucose level of 60 can be mapped to a node of type *laboratory test* and identified by $(Glucose, 60)$. However, in many cases, different values of a unique clinical event convey identical semantic in terms of disease prediction. For instance, various numerical measurements in many laboratory tests are considered the same as long as they fall into one of the normal or abnormal ranges. Therefore, a proper modeling strategy should map clinical events with duplicate diagnostic semantic into the same node as failing to do so can negatively affect the power of the model in capturing similarity of nodes. Having this in mind, following steps are taken for mapping clinical events to nodes. Procedure and diagnosis clinical events are mapped based on the icd-9 coding system [1]. For each laboratory test, its name coupled with a reported flag which can be either normal or abnormal is considered as a unique node. The same strategy is employed when dealing with microbiology tests, where flags can be sensitive, resistant, or intermediate. Moreover,

the age of patients is classified with threshold 15, 30, and 64 based on a statistical analysis of adverse events in different age groups studied by [22]. Finally, gender, ethnicity, and prescription, which are categorical events, are easily mapped by their unique category names.

Symptom Extraction. For extraction of symptoms that are commonly found inside raw-text clinical notes, we employed Autophrase [30] which is a novel phrase mining technique that learns high-quality phrases from a large corpus and allows for incorporating domain-specific knowledge bases for achieving highly domain-relevant results. We feed Autophrase with a pool of clinical phrases that are generated from two main sources: (1) Medical Subject Headings (MeSH) [1] vocabulary treasure which contains 90,000 medical entry terms, and (2) the ICD-10 [2] medical coding database. Quality phrases for symptoms are extracted from MeSH "signs and symptoms" category, code C23 and ICD-10 Chapter XVIII. We also run some final filtering steps on results of Autophrase to drop phrases that include measurements, adverbs, or symbols as they do not contribute to our diagnosis goal.

Having all nodes constructed, their connections are added based on schema in Figure 1 and selected metapaths are discussed in following sections. One of the advantages of HIN is that missing values in EHR only lead to the absence of some links and does not require further management.

3.4 Heterogeneous Network Embedding for Clinical Events

Given the rich clinical information network, learning a latent and low-dimensional embedding of clinical events that can capture their internal relations is greatly beneficial for further analysis tasks. Inspired by the success of skip-gram [25] in learning latent word embeddings from the context of words in a corpus, most of homogeneous network embedding techniques [17, 34] rely on neighbor prediction paradigm. In this approach, given a network $G = (V, E)$, and an embedding function $f : V \rightarrow R^d$ that maps each node to a d dimensional vector, the objective is to maximize the probability of observing neighborhood of a node v, denoted as $N(v)$, conditioned on its representation $f(v)$ [17].

$$\underset{f}{\text{argmax}} \prod_{v \in V} \prod_{c \in N(v)} Pr(c|f(v))$$

where the probability $Pr(c|f(v))$ is defined as a softmax function, normalized with respect to representation of all network nodes.

To exploit the rich structural information of EHR data and enrich semantics of similarity among different nodes, we employ an extension of above paradigm to heterogeneous networks that incorporates variety in node types and metapaths in the definition of node neighborhood and the objective function [6, 13]. In particular, with presence of multiple node types, neighborhood of a node v is defined as $N(v) = \{N_1(v), N_2(v), ..., N_T(v)\}$ where $N_t(v)$ denotes type t neighbors of v and T is the number of node types.

Moreover, having multiple types of paths leaving a node (simple or metapath), the neighbor prediction probability function $Pr(c|f(v))$ should be also conditioned on the type of path used. Specifically, the probability of visiting a neighbor c of a node v under path r with schema $V_1 \rightarrow ... \rightarrow V_l$, is defined as:

$$Pr(c|f(v), r) = \frac{exp(f(c) \cdot f(v))}{\sum_{u \in V_l} exp(f(u) \cdot f(v))}$$

As computation of above probability is very expensive in large networks, negative sampling [25] is employed to achieve following objective function:

$$Pr(c|v, r) = \log \sigma(f(c) \cdot f(v)) + \sum_{1}^{m} \mathbb{E}_{u_l \sim P_l(u_l)} \log \sigma(-f(u_l) \cdot f(v))$$

where m negative sample nodes are drawn based on their node degree and from nodes having the same type as r destination type (V_l). Therefore, a training step randomly samples a path schema r and two nodes v and c connected under r, along with m negative sampled nodes and employs Stochastic Gradient Descent (SGD) to update their embeddings. The tuple (v, c) is sampled based on the normalized number of links under the path r over each node tuples.

Although treatment clinical events should not be directly used in the diagnosis prediction, they can be profoundly beneficial in the unsupervised embedding model for capturing similarity of diagnostic clinical events in terms of consequent treatment. For instance, by including prescription and the metapath $symptom \leftarrow patient \rightarrow prescription$, into the embedding model, it can learn similarity among symptoms that lead to the same prescription. Therefore, for training the unsupervised embedding model, we first select the set of advantageous treatment nodes to be added to the embedding model by evaluating the performance-gain obtained from each or combination of them. Next, among many possible metapaths, we select candidate paths mainly from those that link a diagnostic event to a treatment one through a patient (such as the one above). We also compare candidate metapaths in terms of performance-gain when they are added to the network separately and incrementally and select the best configuration.

3.5 Diagnosis Prediction

When node embeddings are present, the process of diagnosis prediction for a new patient involves construction of the patient's representation based on his clinical events and ranking diagnosis codes according to their dot product similarity to the patient's representation. Figure 2 shows the overview of the prediction flow. Given a patient p, his type t neighborhood ($N_t(p)$) can be summarized into a latent embedding ($f_t(p)$) by averaging its members:

$$f_t(p) = \sum_{n \in N_t(p)} \frac{f(n)}{|N_t(p)|}$$

Having clinical events of p grouped into latent type embeddings ($f_t(p)$), a representation for p can be intuitively achieved by aggregating them, but with different weights for each type (w_t) to capture importance of the type in diagnosis prediction.

$$f(p) = \sum_{t} w_t f_t(p)$$

Finally, a diagnosis d is scored and ranked by a dot product similarity between p and d embeddings: $s(d, p) = f(d) \cdot f(p)$.

[1] https://www.nlm.nih.gov/mesh/
[2] http://www.who.int/classifications/icd/en/

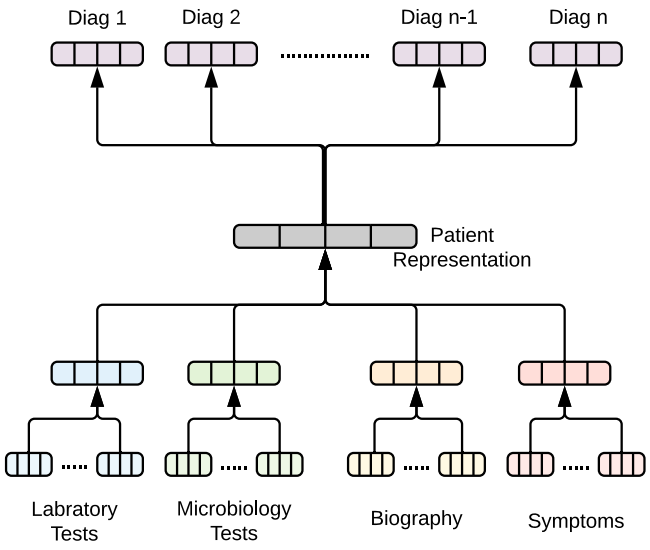

Figure 2: Diagnosis prediction flow.

3.6 Supervised Node Representation Learning for Diagnosis Prediction

The heterogeneous network embedding model discussed in section 3.4, does not have a direct guidance for learning representations that are specifically suitable for disease diagnosis aim and learns a general knowledge of the network. To add such guidance and provide diagnostic knowledge to the model, following [6] we employ the diagnosis prediction flow discussed in section 3.5 as a supervised embedding process and jointly use with the unsupervised model at the time of representation learning to tailor embeddings to disease diagnosis goal.

Recalling computation of prediction score from section 3.5, which is defined for a tuple of diagnosis d and patient p, we have:

$$s(d,p) = f(d) \cdot f(p) = f(d) \sum_t w_t f_t(p) =$$

$$f(d) \sum_t w_t \left(\sum_{n \in N_t(p)} \frac{f(n)}{|N_t(p)|} \right)$$

We can employ a hinge loss ranking objective for the triple (p, d, $\sim d$) to update node embeddings (f) and node type weights (w_t).

$$max(0, -s(d,p) + s(\sim d, p) + \sigma)$$

where d and $\sim d$ are positive and negative sampled diagnosis for p and scores $s(d,p)$ and $s(\sim d, p)$ are calculated for them respectively.

To jointly learn embeddings, objectives of the two supervised and unsupervised models ($\mathbb{Z}_{supervised}$, $\mathbb{Z}_{unsupervised}$) are combined to form the joint objective as:

$$\mathbb{Z}_{joint} = \omega \cdot \mathbb{Z}_{unsupervised} + (1-\omega) \cdot \mathbb{Z}_{supervised} + \lambda \sum_n \|f(n)\|_2^2$$

where $\omega \in [0,1]$ is a pre-defined parameter for tuning importance of either models and a regularization term is added to prevent over-fitting of learned representations.

Therefore, a training step in the joint representation learning model works as follows. We draw one of the embedding models based on $Bernoulli(\omega)$. If the unsupervised model is drawn, its objective function is used on a mini-batch of randomly drawn triples (r, v, c) and m negative samples to update representations. Otherwise, the supervised objective is used for a mini-batch of drawn triples (p, d, $\sim d$) to update type weights (w_t) and representations (f). Negative samples are drawn in both cases from a unigram distribution based on node degree [25].

4 EXPERIMENTS

In this section we evaluate HeteroMed through three sets of experiments. First, it is evaluated under different design configurations. Then its diagnosis prediction performance is compared to various baseline models and finally it is quantitatively evaluated through two case studies.

4.1 Dataset

Experiments of this study are conducted on the publicly available Medical Information Mart for Intensive Care III (MiMIC III) [20] dataset. It contains a comprehensive clinical data for forty thousand patients admitted to the ICU department of BIDMC hospital during 11 years. The MIMIC dataset is organized into 26 tables containing clinical event records for each admission to the ICU and other general information such as definitions of clinical terms. Table 1 lists utilized database tables alongside with main columns used and a short description for each table. In this study, each admission of an adult subject (aged 15 years or older) to the hospital is considered as a sample and called a *patient stay*. Few subjects with multiple ICU stays in a single hospital admission were excluded due to the insufficiency of diagnosis information provided for them in MIMIC.

Following these steps, we obtained a sample set of 46,641 patient stays from which 10,000 were randomly sampled for the test set and 36,641 remaining for the training set. The heterogeneous network was then constructed with the strategy discussed in section 3.3 using our train set. Table 2 lists statistical details for nodes of this network. Furthermore, in addition to 9 length one basic links of the network, 9 other candidate metapaths were selected. Table 3 lists both types of paths with their frequency in the constructed network. As patient node is a central hub in our metapaths, each path is denoted only by its start and end node types. (e.g., lab-symp denotes the *laboratory test*← *patient* → *symptom* metapath).

4.2 Evaluation Strategies and Implementation Details

Disease diagnosis is conducted in two levels in this study. First, exact diagnosis code prediction as a ranking problem and second general disease cohort prediction as a multi-label classification problem which are evaluated with *MAP@k* and *AUROC score* respectively. *MAP@k* is a metric widely employed in information retrieval and reports the mean of average precision at k (*AP@k*) over all test samples. In this study, having a ranked diagnoses list returned by the prediction model, *AP@k* shows the averaged precision over all the positions in the list that the diagnosis is correct and has index less than k.

Table 1: MIMIC tables used in this study.

Table name	Main Columns	Description
patients_icd	gender, DOB, ethnicity	Name and demographic information of patients
procedures_icd	icd9_code	Procedure events such as brain monitoring, tubing, injection
prescriptions	generic_drug_name	Drugs prescribed in each admission
microbiologyevents	spec_itemid, interpretation	Microbiology tests and their sensitivity level; eg. fungi, bacteria
labevents	itemid, flag	Laboratory results and their flag (normal, abnormal); eg. Blood Glucose
Diagnosis_icd	icd9_code	Prescribed diagnosis codes.
noteevents	Category = D̈ischarge Summary¨	Raw text notes recorded by nurses which includes symptoms and other clinical information collected on admission time.

Table 2: Node statistics for the HIN network.

Node Type	abbreviation	Train	Test
Patient stay	pati	36641	10000
Procedures	proc	1673	746
Prescription	pres	6000	3523
Microbiology	micro	212	63
Laboratory	lab	1870	1045
Diagnosis	diag	5605	2745
Symptom	symp	1602	435
Gender	gen	2	2
Age group	age	3	3
ethnicity	eth	40	32

Table 3: Edge statistics for constructed network.

simple links	count	metapaths	count
pati-proc	94,452	lab-diag	1,155,278
pati-pres	757,195	symp-diag	261,861
pati-micro	19,768	lab-proc	341,907
pati-lab	1,948,360	lab-pres	770,297
pati-age	36,641	symp-pres	223,666
pati-diag	292,473	symp-proc	63,424
pati-symp	307,325	lab-symp	214,356
pati-gen	36,641	micro-lab	14,394
pati-eth	35,342	micro-symp	8,696

AUROC is a goodness of binary prediction metric based on different cut-off thresholds on classifier prediction score. Here, AUROC is computed for each of disease cohorts based on the scores computed by our supervised prediction model and baselines for each cohort.

For training HeteroMed and learning node embeddings, a mini-batch of 500 patients has been used at each training step with embedding vector size of 128. The unsupervised embedding model is selected with 4 times higher probability than the supervised model when performing the joint representation learning. Furthermore, each step of unsupervised approach draws 100 negative diagnosis samples for each patient based on the diagnosis node degree.

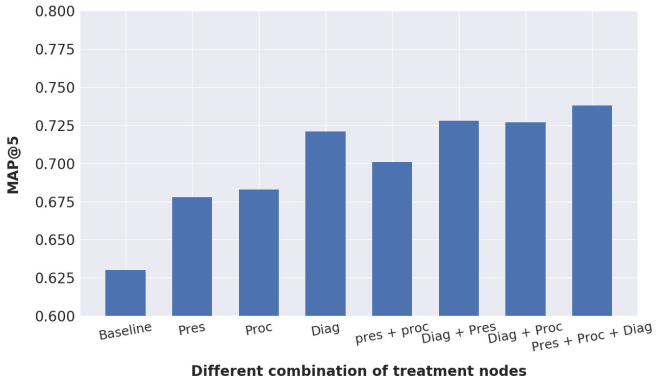

Figure 3: Treatment node selection evaluation.

4.3 Evaluation of Proposed Method

In this section, we demonstrate experimental results of evaluating performance of HeteroMed under different metapaths and node selection configurations.

Treatment Node Selection.
In this part, we evaluate performance-gain obtained, when each or a combination of treatment nodes (proc, pres, diag) are included in the network of unsupervised embedding model. Figure 3 illustrates the comparison to the baseline performance in which the network only contains diagnostic nodes. We can observe that among treatment nodes, "diagnosis" show a great advantage to be added to the unsupervised embedding process. This is partly due to the fact that any improvement in diagnosis node embeddings directly impacts performance of diagnosis prediction. Procedure and prescription nodes also impact the performance in a positive way. This is while they are not included in the prediction step of diagnosis. Results of this experiment confirm the advantage of utilizing the whole set of available information in the unsupervised representations learning process.

Metapath Selection.
In the second part of this experiment, we evaluate performance-gain obtained by using selected metapaths listed in Table 3. The results are elaborated in Figure 4. The blue bars in the figure show the performance for each metapath when added separately to the baseline and are sorted in descending order based on this measure.

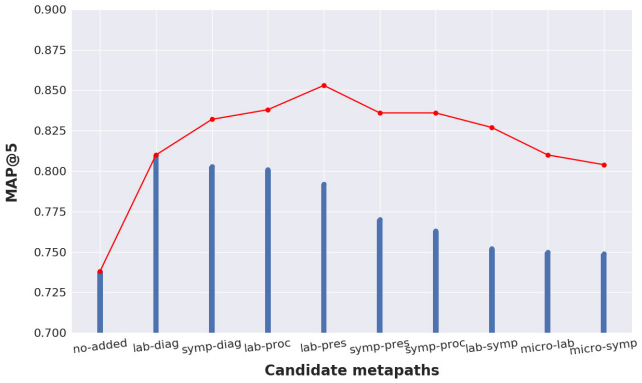

Figure 4: Metapath selection evaluation. Red line denotes additive performance and blue bars denote single path performance.

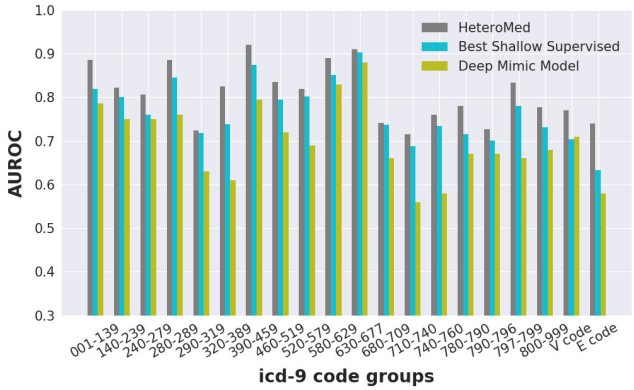

Figure 5: Disease cohort prediction evaluation.

The red line, however, evaluates performance when these paths are accumulated incrementally to the model. Results of this experiment indicate that the combination of 4 first metapaths (lab-diag, symp-diag, lab-symp, lab-pres) provides us with the optimal performance for the disease diagnosis goal. This is while adding more paths leads to a gradual performance drop. This observation further clarifies the significant advantage of metapath-based neighbor sampling rather than the random neighbor sampling used in prior medical domain studies.

Based on these results, the model used in all succeeding experiments employs all treatment nodes and the 4 above-mentioned metapaths in its representation learning process.

4.4 HeteroMed Compared to Other Diagnostic Models

To further assess our model, we compare its diagnosis performance to selected state of the art models in two levels. First, when exact disease codes are to be predicted and second when disease cohorts are desired.

Exact Code Prediction.

In this experiment, we try to rank exact disease codes for a patient stay. In this part, we compare HeteroMed only to embedding based models as the size of diagnosis codes is too large to be predicted by a supervised classifier. The baseline models in this task are:

Med2vec: Med2vec [7] is a multilayer medical embedding neural network which learns embeddings of medical events and visits using an approach inspired by word2vec. We modified the last softmax layer of this model to predict diagnosis codes for the current visit as the model originally predicted disease codes for last or future visits. This method is implemented by Theano python library [2] and representation sizes are chosen to be 100.

Skipgram-embedded: We use word2vec (skip-gram) to learn network node representations similar to the way it is employed in prior studies. In particular, all clinical events associated with an admission are considered as words and are concatenated to form sentences. The window size is set to the maximum length of sentences so that all clinical events (words) in an admission (sentence) can be sampled as neighbors. The method is implemented by the open source python tool, Gensim [29]. Node representations learned are fed into the supervised prediction model (section 3.5) to score diagnoses and rank them.

HeteroMed-embedded: We learn node embeddings by only employing the unsupervised representation learning approach introduced in section 3.4. We use the same set of metapaths as the main model for this aim. As the previous method, the supervised diagnosis model is then used to rank diagnoses.

Table 4: Comparison of HeteroMed model to baselines for exact code prediction.

Model Name	MAP@3	MAP@5	MAP@10
Med2vec	0.75	0.78	0.79
Skipgram-embedded	0.73	0.76	0.77
HeteroMed-embedded	0.78	0.79	0.80
HeteroMed	**0.81**	**0.85**	**0.87**

Disease Cohort Prediction.

Icd-9 diagnosis coding provides 20 code groups that correspond to 20 high level disease cohorts. In this part, we aim to predict all disease cohorts that a patient is diagnosed with. Having multiple diagnosis codes for a patient stay, different groups of diseases may be involved which turns the problem into a multi-label classification. When training our model for cohort prediction, only 20 disease nodes are constructed for the network and each disease code of patient is mapped to one of these nodes. Furthermore, in prediction time, scores for all 20 diagnosis nodes are computed to be evaluated. The baseline models are listed below:

Shallow supervised models: We use feature engineering along with common shallow models, from which Random Forest provided best results. We extracted the same features suggested by [28] but only from tables used to construct our network. We employ Scikit-learn [26] for implementation of the basic models.

Deep Mimic Model: We finally compare our results to the ones from mimic learning model [5] which employs a deep

Table 5: Similarity search results.

Diabetes		Cold		Anemia (lack of blood)	
HeteroMed	skipgram	HeteroMed	skipgram	HeteroMed	skipgram
peripheral neuropathy	**dietary change**	**general pain**	**fever**	**fatigue**	weight loss
sleep apnea	tightness	**fever**	**sick contact**	**malaise**	**allergy reaction to iron**
leg tingling	confusion	**chill**	constipation	**heart palpitation**	penile discharge
urinary frequency	speak difficulty	**sore throat**	**muscle pain**	**itchy skin**	sick contact
ulcers	nausea	swelling	**recent travel**	**bloody stool**	**shortness of breath**
dietary change	**rash**	**allergy**	limb pain	bruising	stuffy nose
burning	fever	**tightness**	urinary changes	**abdominal pain**	**leg tingling**
abdominal pain	mental status change	**sinus congestion**	**cough**	**nausea**	suicidal attempt
thirst	**numbness**	**cough**	stiff neck	**chills**	**abdominal pain**
itchy skin	sleepiness	blurred vision	**runny nose**	**cramps**	**jaundice**

neural network alongside with a Gradient Boosting Model for prediction of icd-9 diagnosis code groups.

Results.
The exact code prediction evaluation is depicted in Table 4. As the results suggest, HeteroMed outperforms all the baseline models in exact diagnosis prediction. The out performance of HeteroMed-embedded model compared to skipgram-embedded model, reveals superiority of relation-aware embedding approach employed in this study to the skip-gram used in conventional clinical models. Furthermore, the Med2vec model outperforms the Skipgram-embedded model although they are both trained based on skip-gram embedding. This can be due to the fact that Med2vec incrementally updates the embeddings with back propagation in its model. However, it sill falls behind HeteroMed that employs relation-aware embedding approach.

Results of the disease cohort prediction are illustrated in Figure 5. We can observe that HeteroMed performance exceeds baseline models in almost all code groups. In general, performance in some groups are lower than the others which generally corresponds to those diagnosis groups that are sparser in the MIMIC dataset.

4.5 Case Studies

In this section, we qualitatively evaluate modeling of EHR data using HeteroMed and validate sensibility of learned clinical event representations. First, we perform a similarity search to find relevant symptoms to three common diseases. We then review results of a sample prediction case. In both experiments, we compare the results to the Skipgram-embedded model introduced in the last section.

Table 5 lists top ten related symptoms and observations to three common clinical conditions: Diabetes, Cold, and Anemia. A dot product similarity has been employed to generate these results. To achieve better vision for comparison, results are validated by a clinical expert and relevant symptoms are shown in bold format. Recognizing the fact that symptoms can have hidden and complex relations to diseases, only directly related symptoms to each condition are considered as relevant.

Results of this experiment confirm the validity of learned representations by our model. Moreover, we can easily observe that HeteroMed ranks relevant symptoms higher than the Skipgram-embedded model and is vividly stronger in understanding relations of symptoms to diseases. One may notice that the intersection among results of two models is small. The large number of symptoms and the fact that a single complication can be described in multiple ways are the main reasons for this observation. For instance, leg tingling, numbness, and peripheral neuropathy can all refer to a similar complication caused by diabetes.

Table 6: Comparison of sample prediction results for a patient and real diagnosis codes. Star sign shows that the predicted code is not present in the ground truth codes.

Ground truth	Category	Skipgram	HeteroMed
4282	Circulatory system	2875	4273
4254	Circulatory system	3970	4282
2875	Blood organs	6841*	4583
4273	Circulatory system	281	2832*
3970	Circulatory system	7217*	2875
5303	Digestive system	427*	4254
4280	Circulatory system	4583	530*
281	Blood organs	4273	260*
4583	Circulatory system	2501*	281

Table 6 shows a sample admission with 9 real diagnosis codes along with the 9 top-ranked predicted codes by each model. Wrong predictions are denoted by a star sign ($*$) on the top right corner. Furthermore, the main category of each disease code in ground truth is specified to provide better understanding of codes. The two methods rank a number of wrong codes in their first 9 predictions. However, the superior performance of HeteroMed is noticeable in two aspects. Firstly, we can observe that HeteroMed is able to detect all disease categories of the ground truth, although not predicting exact diagnosis codes. For instance, HeteroMed ranks the code 530, which is not present in ground truth, in its top 9 predicted codes. However, this code is a more general indication of the code 5303 in the ground truth and both show a digestive system disease with different specificity. This is while the top diagnosis codes ranked by Skipgram-embedded model do not cover this disease

category. Secondly, all the top-ranked codes by HeteroMed are related to disease categories that are present in the ground truth. However, some of the diagnosis codes that are ranked high by Skipgram-embedded model are outside the ground truth disease categories. For instance, the code 7217 which relates to connective tissue diseases is ranked 5 by Skipgram-embedded model, while this disease category is not present in the ground truth.

In general, we can observe that HeteroMed can achieve superior results in major prediction experiments.

5 CONCLUSION AND FUTURE WORK

In this paper, we study the problem of disease diagnosis from a patient's diagnostic records available in EHR data. We propose modeling of clinical events as a heterogeneous information network, HeteroMed, to address shortcomings of previous methods pursuing same goals. Existing studies ignore the rich structure and relations in EHR data when learning representations of clinical events. HeteroMed is capable of capturing informative relations for the diagnosis goal and use the best relation sampling strategy when learning clinical event representations. It also allows for easy handling of missing values and learning embeddings tailored to the disease prediction goal using a joint embedding framework. Result of our study shows that HeteroMed can achieve significantly better results in diagnosis task and finding clinical similarities. This in turn confirms the benefits of employing heterogeneous information network in modeling clinical data.

Future work includes modeling more diverse type of information using heterogeneous network such as timeseries data and joint suggestion of disease and treatment based on available diagnostic information.

ACKNOWLEDGEMENTS

We would like to thank anonymous reviewers for their constructive comments. This work is partially supported by NSF III-1705169, NSF CAREER Award 1741634, and Snapchat gift funds.

REFERENCES

[1] American Medical Association. 2004. *International classification of diseases, 9th revision, clinical modification: physician ICD-9-CM, 2005: volumes 1 and 2, color-coded, illustrated*. Vol. 1. Amer Medical Assn.

[2] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: A CPU and GPU math compiler in Python. In *Proc. 9th Python in Science Conf.* 1–7.

[3] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. 2010. Secondary use of EHR: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics* 2010 (2010), 1.

[4] Shiyu Chang, Wei Han, Jiliang Tang, Guo-Jun Qi, Charu C Aggarwal, and Thomas S Huang. 2015. Heterogeneous network embedding via deep architectures. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 119–128.

[5] Zhengping Che and Yan Liu. 2017. Deep Learning Solutions to Computational Phenotyping in Health Care. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*. IEEE, 1100–1109.

[6] Ting Chen and Yizhou Sun. 2017. Task-guided and path-augmented heterogeneous network embedding for author identification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 295–304.

[7] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1495–1504.

[8] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 787–795.

[9] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*. 3504–3512.

[10] Youngduck Choi, Chill Yi-I Chiu, and David Sontag. 2016. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings* 2016 (2016), 41.

[11] Youngduck Choi, Chill Yi-I Chiu, and David Sontag. 2016. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings* 2016 (2016), 41.

[12] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. 2014. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. ACM, 1819–1822.

[13] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 135–144.

[14] Wael Farhan, Zhimu Wang, Yingxiang Huang, Shuang Wang, Fei Wang, and Xiaoqian Jiang. 2016. A predictive model for medical events based on contextual embedding of temporal sequences. *JMIR medical informatics* 4, 4 (2016).

[15] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, Russ B Altman, and Roded Sharan. 2013. A method for inferring medical diagnoses from patient similarities. *BMC medicine* 11, 1 (2013), 194.

[16] Mark L Graber, Nancy Franklin, and Ruthanna Gordon. 2005. Diagnostic error in internal medicine. *Archives of internal medicine* 165, 13 (2005), 1493–1499.

[17] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 855–864.

[18] Jiawei Han, Yizhou Sun, Xifeng Yan, and Philip S Yu. 2010. Mining knowledge from databases: an information network analysis approach. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 1251–1252.

[19] Shifu Hou, Yanfang Ye, Yangqiu Song, and Melih Abdulhayoglu. 2017. Hindroid: An intelligent android malware detection system based on structured heterogeneous information network. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1507–1515.

[20] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3 (2016), 160035.

[21] Rong-Ho Lin. 2009. An intelligent model for liver disease diagnosis. *Artificial Intelligence in Medicine* 47, 1 (2009), 53–62.

[22] Jake Luo, Christina Eldredge, Chi C Cho, and Ron A Cisler. 2016. Population Analysis of Adverse Events in Different Age Groups Using Big Clinical Trials Data. *JMIR medical informatics* 4, 4 (2016).

[23] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1903–1911.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[26] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[27] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 701–710.

[28] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. 2017. Benchmark of Deep Learning Models on Large Healthcare MIMIC Datasets. *arXiv preprint arXiv:1710.08531* (2017).

[29] Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

[30] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2017. Automated phrase mining from massive text corpora. *arXiv preprint arXiv:1702.04457* (2017).

[31] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. 2017. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge*

and Data Engineering 29, 1 (2017), 17–37.

[32] Chuan Shi, Zhiqiang Zhang, Ping Luo, Philip S Yu, Yading Yue, and Bin Wu. 2015. Semantic path based personalized recommendation on weighted heterogeneous information networks. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 453–462.

[33] Jyoti Soni, Ujma Ansari, Dipesh Sharma, and Sunita Soni. 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications* 17, 8 (2011), 43–48.

[34] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1067–1077.

[35] Cheng-Hsiung Weng, Tony Cheng-Kui Huang, and Ruo-Ping Han. 2016. Disease prediction with different types of neural network classifiers. *Telematics and Informatics* 33, 2 (2016), 277–292.