Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 19(6): 5832–5849.
DOI: 10.3934/mbe.2022273
Received: 18 December 2021
Revised: 16 March 2022
Accepted: 22 March 2022
Published: 06 April 2022

Research article

PercolationDF: A percolation-based medical diagnosis framework

Jingchi Jiang[1], Xuehui Yu[2], Yi Lin[2] and Yi Guan[2,*]

[1] The Artificial Intelligence Institute, Harbin Institute of Technology, Harbin, China
[2] The Faculty of Computing, Harbin Institute of Technology, Harbin, China

* **Correspondence:** Email: guanyi@hit.edu.cn.

Abstract: *Goal:* With the continuing shortage and unequal distribution of medical resources, our objective is to develop a general diagnosis framework that utilizes a smaller amount of electronic medical records (EMRs) to alleviate the problem that the data volume requirement of prevailing models is too vast for medical institutions to afford. *Methods:* The framework proposed contains network construction, network expansion, and disease diagnosis methods. In the first two stages above, the knowledge extracted from EMRs is utilized to build and expense an EMR-based medical knowledge network (EMKN) to model and represent the medical knowledge. Then, percolation theory is modified to diagnose EMKN. *Result:* Facing the lack of data, our framework outperforms naïve Bayes networks, neural networks and logistic regression, especially in the top-10 recall. Out of 207 test cases, 51.7% achieved 100% in the top-10 recall, 21% better than what was achieved in one of our previous studies. *Conclusion:* The experimental results show that the proposed framework may be useful for medical knowledge representation and diagnosis. The framework effectively alleviates the lack of data volume by inferring the knowledge modeled in EMKN. *Significance:* The proposed framework not only has applications for diagnosis but also may be extended to other domains to represent and model the knowledge and inference on the representation.

Keywords: complex networks; knowledge representation; medical diagnosis; percolation theory

1. Introduction

Intelligent medical decision support systems (IMDSSs) have become increasingly popular in recent years. These systems can provide clinicians and patients with computer-generated clinical knowledge and give patient-related treatment recommendations by collecting the medical experiences of previous patients [1,2], thus, effectively alleviating the problems of shortage and unequal distribution of medical resources.

To build effective IMDSSs, it is necessary to acquire a vast amount of reliable and high-quality medical knowledge. Most clinical medical knowledge comes from medical text records [3]. An electronic medical record (EMR) is the storage of health data and medical history of patients in an electronic format and, thus, is a rich resource for clinical research [4,5]. With the rapidly growing EMR data, several useful methods based on machine learning techniques have been developed to assist clinicians and patients.

Machine learning techniques provide important technical support in IMDSSs and can process an amount of medical knowledge that exceeds the capacity of the human brain. Under the paradigm of supervised learning from the collective experience of many patients, the system provides diagnosis, management decisions and therapy for other patients [6]. However, even in the smallest department, the diversity of symptoms and diseases makes the amount of data needed for machine learning models too large to be borne by medical institutions.

The “big P, small N” problem is a vivid description for this situation, which means an insufficient number of samples for training considering the high-dimensional features possessed by each sample [7]. From the descriptive statistics, there were a total of 5840 symptoms and 1066 diagnoses in only 992 EMRs. Symptoms form the feature set, which includes 5840 dimensions, and diagnoses form the label set, which includes 1066 labels but only 992 samples. On average, the number of samples for each label is less than 1, and the number of feature dimensions is very high (5840). In addition, a key characteristic that can be obtained in our data is that the labels have a very strong causality only with respect to certain feature dimensions, and the certain features for different labels are in different dimensions.

Machine learning techniques are usually unable to obtain good knowledge from high-dimensional features with statistical methods. Few-shot learning can turn limited supervised experience into useful prior knowledge. This style of learning mimics the human ability to acquire knowledge from a few examples through generalization and analogy [8]. According to this idea, extracting the knowledge from the “little” data into a suitable model to represent the knowledge, and then diagnosis with the knowledge learned before, may be a better and efficient method.

Networks have been widely researched to model medical knowledge [9–13]. In a previous study, Zhao et al. [14] built a medical knowledge network based on EMRs (EMKN) to represent medical knowledge. Their construction rules were modified, and a modified EMKN was built as the knowledge representation in our study. Afterward, percolation theory was incorporated into our diagnosis method making it work efficiently on an evolving graph.

In our study, a diagnosis framework that can provide the scores of diseases in all different departments according to patient symptoms was developed. In this paper, a percolation-based diagnosis framework (PercolationDF) containing three parts, network construction, network expansion and diagnosis methods, is proposed. Initially, a base EMKN was built utilizing the network construction method which models the process of knowledge learning. Afterward, the EMKN was supplemented with the network expansion method, which models the process of acquiring new knowledge that has not been learnt before. Finally, our diagnosis method works on EMKN for diagnosis.

The forllowing are three contributions of this paper:

1) A diagnosis method incorporating percolation theory, which can accumulate clinical evidence with the rich support of a graph structure and is suitable for operating on evolving graphs is proposed.
2) The EMKN utilized in one of our previous studies is modified in to further clarify the causality, and

the EMKN models the knowledge based on medical data.
3) PercolationDF is suitable for studying few-shot learning, in which the labels have a very strong causality only with respect to certain feature dimensions, and the certain features for different labels are in different dimensions.

The remainder of this paper is structured as follows: in Section 2, some methods and characteristics related to our study are discussed. Then, in Section 3, the details of the PercolationDF, including acquisition of medical knowledge from EMRs to construct a graph-based EMKN and gathering new knowledge from new EMRs to expand the graph structure, are presented. The introduction of percolation theory and description of how this theory helps to accumulate clinical evidence with the rich support of a graph structure are also given. In Sections 4 and 5, our framework is evaluated by using actual EMRs, and the results are discussed. The conclusion and the focus of future studies are presented in Section 错误!未找到引用源。.

2. Related work

2.1. Machine learning methods for intelligent disease diagnosis

In clinical decisions, the application of machine learning techniques to diagnose diseases has attracted public interest. Alizadehsani et al. [15] applied sequential minimal optimization-based algorithms to diagnose coronary artery disease and achieved the best accuracy of 94.08%. Rau et al. [16] built an artificial neural network and logistic regression prediction framework to predict the development of liver cancer in type 2 diabetes mellitus patients; the best results of sensitivity and specificity were 0.757 and 0.755, respectively. In these studies, the corresponding features of patients were manually selected, and; then, the diagnostic problems were converted into a classification study. Therefore, the performance of the framework depends on the quality of the selected features.

The rapid growth of EMR data has prompted the use of deep learning frameworks, which have demonstrated state-of-the-art performance in diagnostics [17–20]. This takes advantage of avoiding manual feature extraction; however, the process of training optimal deep learning frameworks requires a large data volume that most medical institutions cannot afford, which limits these frameworks to perform sufficiently [21].

2.2. Medical knowledge representation with network

Medical networks, which combine systems biology and network science, aim to study the causes of all human diseases, and it is very important to understand the relationship between abnormal examinations (vital sign parameters go beyond the normal scope) and diseases. The main idea is that the human body is a highly complex physiological system, and there are a series of interactions, restrictions, promotions and stimulations between different physiological states. Barabási et al. [10] used a network to model a human physiological system and explore series of interactions, restrictions, promotions, and stimulations between different physiological states. Many medical networks have also been established to study the interrelationships of diseases [11]. César A. Hidalgo et al. [12] built a disease phenotypes network using comorbidity patterns and aimed to capture disease progression patterns and found that patients tend to develop diseases in the subnetwork. In recent years, the network has also been widely adopted in modeling medical knowledge. Jiang et al. [22] proposed a medical knowledge network (MKN) for medical diagnosis, and it can facilitate studies of intelligent diagnosis comparing a Markov logic network and the logistic regression algorithm on an EMR database. Zhao et al. [14] built an EMKN and made good use of it for diagnosis. However, they ignored that in different sections of one record, medical entities with the same word have different meanings. This causes confusion regarding the causality between symptoms and diseases, making some edges redundant and increasing the network complexity.

Established in their work, the network is also utilized to model and represent our medical knowledge from EMRs in our study. The EMKN construction rules presented by Zhao et al. [14] were modified to further clarify the causality and an expansion method was proposed to supplement it.

2.3. Graph-based inference methods

Probabilistic graphic models (PGMs) are a good machine learning method to Modela joint probability distribution via a graph and have been used to diagnose or predict diseases. D.E. Heckerman et al. [23] developed Pathfinder, the first expert system for hematopathology diagnosis based on a Bayesian network. Klann et al. [24] implemented an adaptive recommendation system to provide a treatment menu based on the previous order. Additionally, Flores et al. [25] applied the Bayesian network to the study of heart failure. Their models perform well and confirm that the Bayesian network is a good model for portraying the complex interactions between medical entities. Unfortunately, parameter estimation of the Bayesian network becomes almost impossible with the increase in network scale simply because of the computational complexity of the structure [26]. Thus, the PGM can only be used in a single, specific field, and the parameters of the PGM need to be retrained after supplementing the graph with new medical data.

Graph convolutional neural networks (GCNs) [27] are a great deep learning method inference on a graph structure and are widely used for node embedding, node representation, and node semisupervised prediction. Wee et al. [28] employed a spectral GCN to identify Alzheimer’s disease. The authors trained on a sizable Caucasian dataset from the ADNI cohort [29] and evaluated it on an Asian population to demonstrate the generalization of the classifiers learned. GCNs have received much interest because they use implicit information in biological systems, with interactive nodes connected by edges whose weights might be either temporal correlations or anatomical junctions, according to David Ahmedt-Aristizabal et al. [30]. However, its parameters also need to be retrained when the graph changes, and it has not been widely researched in medical diagnosis.

3. Methods

3.1. Network construction and network expansion

Apart from the insufficient number of samples for training considering the high-dimensional features possessed by each sample, another key characteristic that can be obtained in our data is that the labels have a very strong causality only with respect to certain feature dimensions, and the certain features for different labels are in different dimensions.

In a previous EMKN, Zhao et al. [14] established edges between diseases (which may not be diagnosed) and symptoms when they occurred together in the same record. The authors ignored that in different sections of one record, medical entities with the same word have different meanings and that a disease not located in the diagnosis section may be a symptom for another disease. This causes confusion regarding causality between symptoms and diseases, making some edges redundant and increasing the complexity of the network. To keep the causalities from symptoms to diagnosis which shows a high probability of a diagnosis caused by certain symptoms, the construction rules of previous EMKN were modified in our EMKN.

In the modified EMKN, the diagnosis nodes are formed from entities in the diagnosis section of each record, the symptom nodes are formed from entities in all other sections of each record, and every direct edge starts from a symptom node to a diagnosis node when the two corresponding entities co-occur in one record. As a result, every edge starts from symptom to diagnosis and contains the causality that denotes a disease probably caused by a symptom. For each diagnosis, all the relevant symptoms can be easily found through the edges. For each symptom, all the diseases that probably developed from them can be found in the same way.

To build and supplement an EMKN obeying the rules above, Algorithms 1 and 2 are adopted separately and some slices of the network evolving process are shown in Figure 1.

EMR_set is defined as a set of EMRs:

$$EMR\_set = \{EMR_i(\mathbb{D}_i, \mathbb{S}_i)\}$$

where $EMR_i(\mathbb{D}_i, \mathbb{S}_i)$ is the i-th EMR in EMR_set. $\mathbb{D}_i$ is a diagnosis set of EMR_i, and $\mathbb{S}_i$ is a symptom set of EMR_i.

Algorithm 1: Network construction algorithm

```
Input: $EMR\_set$
Output: EMKN $\mathcal{G}(\mathcal{V}, \mathcal{E})$
*** Network construction process ***
initialize the node set $\mathcal{V}$ and edge set $\mathcal{E}$
for $EMR(\mathbb{D}, \mathbb{S})$ in $EMR\_set$ do:
  for $v$ in $\mathbb{D}$ do:
    $\mathcal{V} \leftarrow \mathcal{V} + node(v)$
    for $\omega$ in $\mathbb{S}$ do:
      $\mathcal{V} \leftarrow \mathcal{V} + node(\omega)$
      $\mathcal{E} \leftarrow \mathcal{E} + edge(v, \omega)$
    end for
  end for
end for
*** Obtain EMKN ***
return $\mathcal{G}(\mathcal{V}, \mathcal{E})$
```

Algorithm 2: Network expansion algorithm

```
**Input:** EMKN $\mathcal{G}(\mathcal{V},\mathcal{E})$;
        $EMR(\mathbb{D},\mathbb{S}), \mathbb{D}\ is\ a\ \ diagnosis\ set, \mathbb{S}\ is\ a\ symptom\ set$
**Output:** EMKN $\mathcal{G}(\mathcal{V},\mathcal{E})$
*** Update diagnosis nodes ***
for $\omega$ in $\mathbb{S}$ do:
  if $\omega$ not in $\mathcal{V}$:
    $\mathcal{V} \leftarrow \mathcal{V} + node(\omega)$
end for
*** Update symptom nodes and update edges ***
for $v$ in $\mathbb{D}$ do:
  if $v$ not in $\mathcal{V}$:
    $\mathcal{V} \leftarrow \mathcal{V} + node(v)$
  for $\omega$ in $\mathbb{S}$ do:
    $\mathcal{E} \leftarrow \mathcal{E} + edge(v,\omega)$
  end for
end for
*** Obtain graph structure ***
return $\mathcal{G}(\mathcal{V},\mathcal{E})$
```

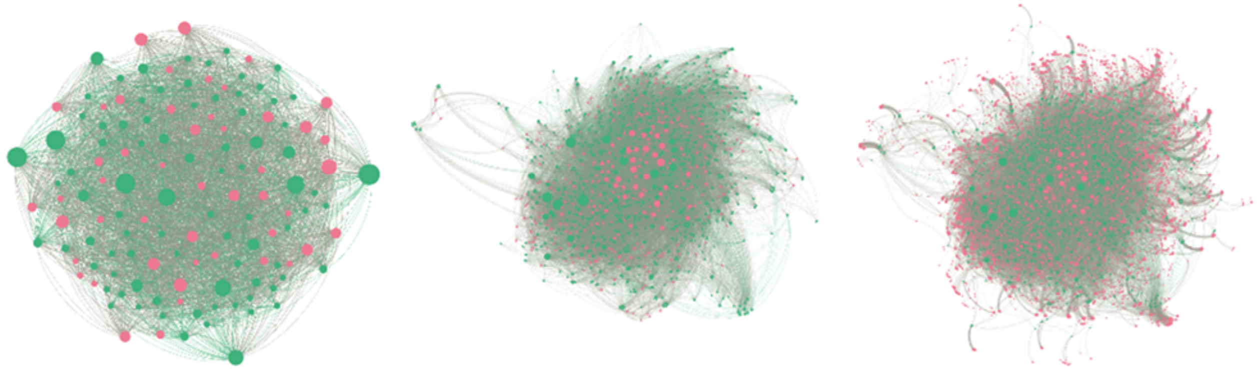


Figure 1. Three slices of the network evolving process visualized by Gephi [31].

Green nodes denote symptoms and pink nodes denote diagnoses. With the graph evolving, it can be easily realized that the number of nodes and edges is increasing. The left figure depicts a slice containing 132 (1.99%) nodes and 3380 (6.13%) edges. The middle figure depicts a slice containing 1174 (17.67%) nodes and 3380 edges. The right figure depicts the end slice which contains all 6645 nodes and 77,179 edges.

3.2. Disease diagnosis

A diagnosis method incorporating percolation theory is proposed and is suitable for operating on evolving graphs to obtain scores of diagnoses in all different departments according to the patient's symptoms.

In the complex network [9] domain, percolation theory, which is widely researched to model the influence spreading of real-world systems, has also been used to model the spread of disease [32,33]. However, the application of this theory to diagnostics has not been comprehensively studied. In our study, the percolation process was modified into clinical evidence percolation in an EMKN and utilized for diagnosis.

Percolation theory is a starting point for resolving the question "Suppose a large porous stone is immersed in water. What is the probability that one point of the stone is wet?", according to Grimmett [34]. When a porous stone is thrown into water, the inner passageways of the stone are *open* with a probability p or *closed* otherwise because of the fluid pressure in the pores. Whether water will fill one point of the stone is considered, and the assumptions are as follows:

1) The greater the value of p, which is the probability of the inner passageway being *open*, the more easily the point of the stone becomes wet.

2) The greater the number of passageways, which connect with the point, the more easily the point of the stone becomes wet.

To make use of it in the network, percolation processes are mostly simple "nodes and edges" percolation that describes the process of the influence spreading in a large graph among different subgraphs, and nodes or edges on this graph are designated as either *"occupied"* or *"unoccupied"* [9]. The state *"occupied"* or *"unoccupied"* of one node means the node has received or not received the influence spread from its edge. The state *"occupied"* or *"unoccupied"* of one edge means that the influence from one of its nodes has crossed through this edge or has not crossed through, as the latter suggests.

In traditional percolation frameworks [错误!未找到引用源。], the influence of subnetwork X, which denotes one part of the whole system, can be quantified as a W_X function:

$$W_X(\beta) = \sum_{k=0}^{\infty} P_X(k) \sum_{j=0}^{k} \binom{k}{j} \beta^j (1-\beta)^{k-j} r_X(j,k) \tag{1}$$

where β represents the ratio of activated edges between network X and its opposite network, which denotes another part of the whole system. $P_X(k)$ is the degree distribution of edges in network X, and $r_X(j,k)$ represents the probability of nodes with degree k becoming activated when the j neighbors were activated W_X denotes the ratio of activated nodes in network X and is a real number in the range [0, 1]. It fits the process of network X receiving influence from its opposite network, which has already been influenced with probability.

This function is based on a random graph, and W_X represents the severity of influence obtained from the opposite network; however, this function is not suitable for a diagnosis process, which is illustrated as follows:

1) The influence severity for only the whole subnetwork X is available, but the influence severity for every node in network X is needed, but not available.

2) It is incorrect to assume that nodes with the same degree k receive the same influence severity. In fact, different diagnoses are relevant to different symptoms, and some symptoms can be associated with different diagnoses.

The ratio of activated nodes and edges in network X are real numbers in the range [0, 1], and which nodes and edges are activated is uncertain. However, it is a certain event that a patient has a symptom or not.

Considering the above, Algorithm 3 is proposed to fit a diagnosis process on an EMKN. State is defined as a parameter for every symptom node with a real number 1, or 0, separately to denote occupied or unoccupied, meaning an associated symptom occurs within a patient or not.

Initially, the state of all symptom nodes, whose associated symptoms are presented by the patient are set as 1, and others are set as 0. Then, the percolation process works and medical evidence spreads. All the diagnosis nodes are influenced by their neighboring symptom nodes. This process is calculated by $PERCOLATE$, which will be introduced in Section 3.3. Next, diagnosis list $\mathcal{D}$ is reranked by d_v, which is the score of disease v. Finally, diagnosis list $\mathcal{D}$ returns, and the higher the score d_v is, the more likely the patient is to have the disease v.

```
Algorithm 3: Disease diagnosis algorithm
Input: EMKN $\mathcal{G}(\mathcal{V},\mathcal{E})$;
        symptom set $\mathcal{S}$ of one patient;
        percolation function $PERCOLATE$
Output: Diagnosis list $\mathcal{D}$
*** Initialization process ***
1 $\text{State}_v \leftarrow 0$ , $\forall v \in \mathcal{V}$
2 $\text{State}_v \leftarrow 1$ , $\forall v \in \mathcal{S}$
*** percolation process ***
3 for $v$ in all the diagnosis nodes do:
4       $d_v \leftarrow PERCOLATE(\{\text{State}_s, \forall s \in Ne(v)\})$, $Ne(v)$ denotes the neighbors of $v$.
5 end for
*** Obtain diagnoses' score ***
6 $\mathcal{D} \leftarrow$ Sorted $\{v, \forall v$ is a diagnosis node$\}$ by the $d_v$
7 return $\mathcal{D}$
```

3.3. C. PERCOLATE process

The influence spreading among nodes through edges is a good idea that can be used to model the process of diagnosis. In a diagnosis process, a patient's symptoms can be regarded as an occupied node, and the influence spreads to the disease. Applying this idea to our study, some modifications for *PERCOLATE* are used to calculate how the diagnosis nodes were influenced by their symptom neighbors.

The score of a disease node depends only on the state of its symptom neighbors. When calculating the score of the diagnosis node, *PERCOLATE* is faded with its neighbors' states. The formula for calculation is as follows:

$$d_v = PERCOLATE(\{\text{State}_s, \forall s \in Ne(v)\}) = \sum_{s \in Ne(v)} R(State_s) \quad (2)$$

where d_v denotes the score of disease v; $Ne(v)$ is the neighbor of node v, which is a symptom node set; $State_s$ represents the state of node s, which takes a value of 0 or 1; and $R(State_s)$ represents a percolation process according to the state of node s and is determined by the task requirements. The score $R(State_s)$ measures the causal effect of feature s on diagnosis v. While the score $R(State_s)$ of feature s is large enough, the evidence s can produce causality support for the diagnosis v

according to (2).

Here, an assumption that each symptom has the same influence on a disease is presented; $R(\cdot)$ can be calculated as:

$$R(State_s) = \begin{cases} 1, & State_s = 1 \\ 0, & State_s = 0 \end{cases} \tag{3}$$

The medical evidence percolation can be rewritten in a compact form as:

$$\mathcal{D} = PERCOLATE(\mathcal{S}, \mathcal{G}(\mathcal{V}, \mathcal{E})) \tag{4}$$

Here, $PERCOLATE$ is a *global percolation function*, which is faded by an EMKN $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and a symptom set $\mathcal{S}$ of a patient. Equation (4) is a compact form of (3), which is an exact, detailed implementation. After the improvement of percolation theory, the three inapplicabilities above have been resolved:

1) The influence severity of every symptom node is available after initialization, and for every diagnosis node, it is available after the percolation process.

2) Each diagnosis node obtains different influence severities according to the severity of its neighbors.

3) The activation of all nodes and edges is certain according to the symptoms a patient has.

4. Experiments and discussion

4.1. Preliminaries on data

A total of 992 Chinese EMRs were used in our study and retrieved from the Second Affiliated Hospital of Harbin Medical University, which contained 887 individual patients. The private information of all the patients was removed and usage rights for these records wereobtained.[a]

The next task required a great quantity of manually labeled data. First, the *medical concept annotation guideline* and *assertion annotation guideline* published by Informatics for Integrating Biology and the Bedside (i2b2) [36,37] were referenced to layout guidelines for manual annotation under the guidance of medical professionals [38]. In the manual annotation process, medical entities were classified into five classes: *disease*, *complaint symptom*, *test*, *test result*, and *treatment*. In this study, *disease* in the 'Assessment and Diagnosis' part of the record is interpreted as *diagnosis,* and medical entities (including *disease*, *complaint symptom*, *test*, *test result*, and *treatment*) in other parts of the record are interpreted as *symptoms* [39].

Figure A1 in the Appendix shows one discharge note from our records corpus. A standard discharge note contains the following six sections: assessment and diagnosis, admission situation, treatment, discharge situation, treatment effect, and discharge order. In addition, the corresponding annotations about this progress note are listed in Table A1.

4.2. Experiment setup

Out of the total gathered EMRs, 500 were randomly selected to build a basis EMKN based on the network construction algorithm, and randomly selected EMRs to expand the EMKN based on the network expansion algorithm. After selecting 785 EMRs in total, the graph structure no longer changes,

[a] https://github.com/WILAB-HIT/Resources.

so the 785 EMRs are used to build the EMKN, and the remaining 207 EMRs of the sample data are used to test and evaluate the performance of our framework.

The other three models as baselines implemented with default settings in the scikit-learn library use naïve Bayes, neural networks[b], and logistic regression, which are widely used in many of the recent similar studies because of the simplicity, effectiveness, and robustness of these methods [35,40,41].

4.3. Evaluation metrics

Many standard evaluation measures, such as the precision-recall curve, receiver operating characteristic (ROC) curve, and area under the curve (AUC), are adopted in diagnostic support systems for specific diseases, but these measures are unsuitable for our task because these curves are meaningful only when the number of positive instances is high enough (greater than 10, for example) [42]. In extreme cases, one EMR may be only one diagnosed disease; that is, that instance has only one positive instance, and the recall rate (or sensitivity on the ROC curve) would be either 0 or 1. Therefore, traditional evaluation measures are less useful for evaluation or comparison. Our task involves returning diseases from a fixed entity set; this task seems to be an information retrieval (IR) task. Therefore, four IR evaluation measures are adopted to evaluate and compare the performance of a single test instance: top-*k* recall ($R@k$), top-*k* precision ($P@k$), *R-precision* and average precision (*AP*).

Top-k recall ($R@k$). $R@k$ is defined as follows:

$$R@k = \frac{\#\ of\ true\ positive\ returned\ in\ top\ k\ items}{\#\ of\ positive\ in\ records} \tag{5}$$

This metric is consistent with the differential diagnosis framework, where the machine suggests *k* possible items and we measure the fraction of true items that are correctly returned. $R@k$ mimics the behavior of doctors conducting clinical diagnoses, where doctors list the most likely diagnoses according to the patient's symptoms. Therefore, a machine learning method with a high $R@k$ is equivalent to a doctor with an effective diagnostic skill, thus making $R@k$ a powerful tool for evaluating the performance of the frameworks in addressing our problem. This measure was also utilized by Choi et al. [35].

Top-k precision ($P@k$). $P@k$ is defined as follows:

$$P@k = \frac{\#\ of\ true\ positive\ returned\ in\ top\ k\ items}{k} \tag{6}$$

This measure is not intended for any value of k, which is often set to a smaller number. If there is only one diagnosis in an EMR, then there is only one positive *Y* in the test sample and $P@10$ will not be higher than 0.1. For example, Miotto et al. [43] assigned k with small values (1, 3, 5); for each value of k, the authors compared their framework with a theoretical upper bound, which reports the best results possible (i.e., all the diseases are correctly assigned to each patient).

Alternatively, *R-precision* is defined as the precision after *R* positive diagnoses are returned; *R* is the exact number of positive diagnoses [44]. We assigned k as the exact number of positive diagnoses

[b] A three-layer feed-forward neural network (with 100 hidden layers having a dimensionality that is half that of the input layer and with rectified linear units (ReLUs) as the activation function) and the scikit-learn (formerly scikits learn and known as sklearn) library are utilized. The structure and the activation function are fine-tuned to improve performance.

in the evaluated test EMR so that the theoretical upper bound would always be 1.

Average precision (AP). AP [45,46] is the mean of the precision scores obtained after each true positive diagnosis is returned and is defined as:

$$AP = \frac{1}{N}\sum_{r=1}^{R} P@r \times \delta(r) \quad (7)$$

where N denotes the number of true positive diagnoses in an EMR, R represents the number of all diagnoses, and $P@r$ represents the precision of the top r returned items. The value of indicator function $\delta(r)$ is 1 if the r-th retrieval item is a true positive diagnosis; otherwise, the value is 0. The calculation of AP can be simplified to:

$$AP = \frac{1}{N}\sum_{i=1}^{N} \frac{i}{position(i)} \quad (8)$$

where N denotes the number of true positive diagnoses in an EMR and $position(i)$ represents the position of the i-th true positive diagnosis in returned items. The ideal result is that true positive diagnoses are all returned and ranked at the top of the returned items; in this case, AP = 1. This measure has been demonstrated to be one of the most stable and discriminating measures for evaluating IR systems [44,45].

4.4. Results

The symptom-disease network is shown in Figure 2. To further describe the problem of insufficient data faced in our study, the frequency of all diagnoses was counted, as shown in Figure 3. Algorithm 3 is operated on EMKN to diagnose diseases, and compare its performance with naïve Bayes, neural networks, and logistic regression. Figure 4 shows the distribution of $R@10$, $P@1$, *R-precision* and *AP* in our testing set, where Figure 5 shows the cumulative distribution of $R@10$ and *AP*.

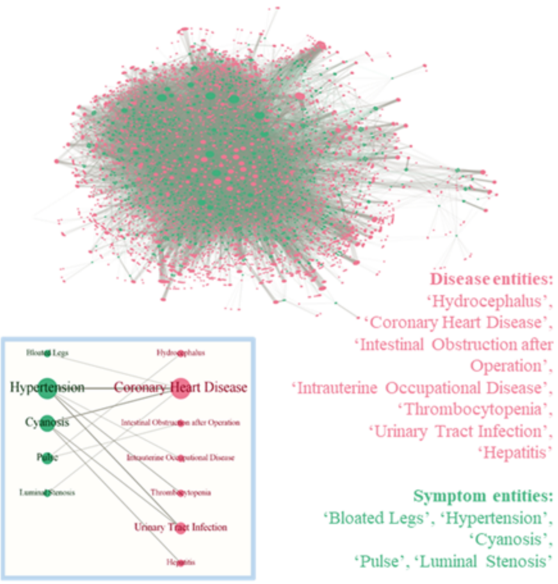

Bloated Legs
Hydrocephalus
Hypertension
Coronary Heart Disease
Cyanosis
Intestinal Obstruction after Operation
Pulse
Intrauterine Occupational Disease
Luminal Stenosis
Thrombocytopenia
Urinary Tract Infection
Hepatitis
Disease entities:
'Hydrocephalus',
'Coronary Heart Disease',
'Intestinal Obstruction after Operation',
'Intrauterine Occupational Disease',
'Thrombocytopenia',
'Urinary Tract Infection',
'Hepatitis'
Symptom entities:
'Bloated Legs', 'Hypertension',
'Cyanosis',
'Pulse', 'Luminal Stenosis'

Figure 2. The symptom-disease network visualized by Gephi [31].

The main figure depicts a visualization of the symptom-disease network, which is generated from 785 records. This network is a bipartite network consisting of only symptom and disease entities. The pink nodes represent disease entities, and the green nodes represent symptom entities. The size of the nodes indicates their degree, and the thickness of the edge indicates the frequency of this relationship. The bottom left corner shows a tiny subgraph of the symptom-disease network, and the label of each node is its unique name. The bottom right corner represents entities on the top right corner; for example, "Hydrocephalus" is a disease entity colored pink, and "Cyanosis" is a symptom entity colored green.

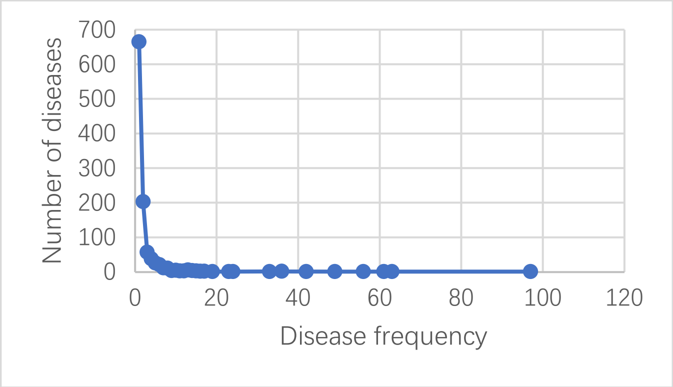


Figure 3. Number of diseases with different frequencies. Of all the disease nodes, 665 occurred only once, and 203 occurred only twice.

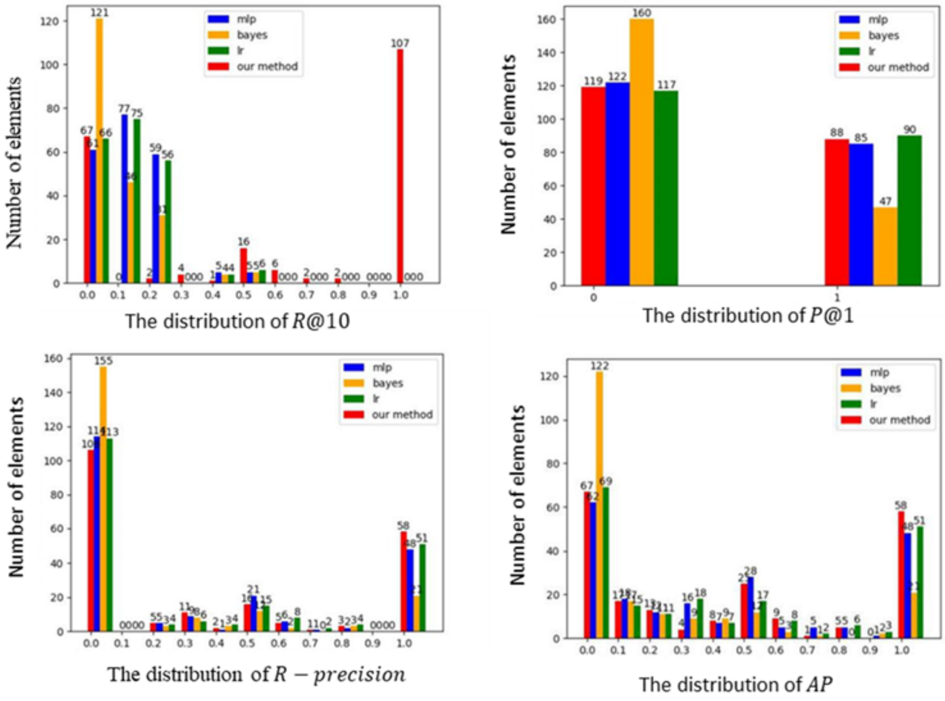


Figure 4. Distribution of R@10, P@1, R-precision and AP. The y-axis displays the number of records.

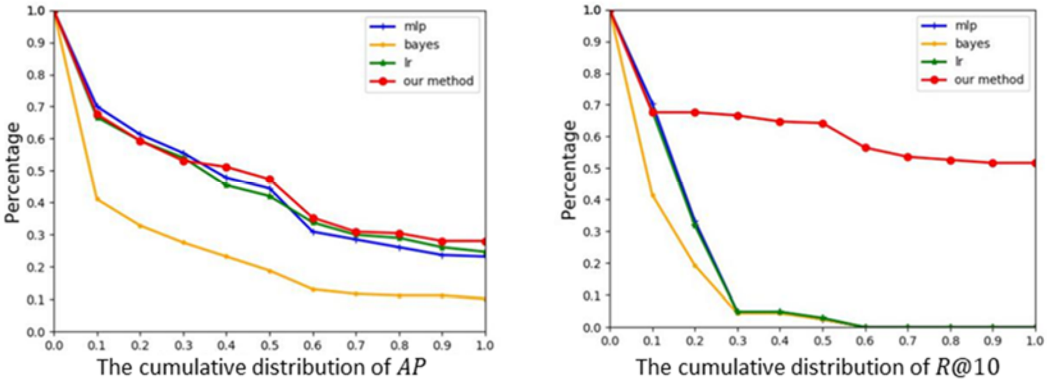


Figure 5. Cumulative distribution of R@10 and AP. The test cases with evaluation measures that exceed the number in the x-axis are counted, and the y-axis displays the corresponding percentage of the count.

5. Discussion

In the EMKN, there are 5840 nodes in the symptom set, 1066 nodes in the diagnosis set and 55,415 edges. This is a bipartite interdependent network structure concluding a symptoms subgraph (symptom network) and a diagnosis subgraph (diagnosis network).

Of all the disease nodes, 665 occurred only once, and 203 occurred only twice. This means that in our framework, over 90% of the diagnoses were met less than 3 times. In this situation, the heavily unequal distribution of data samples will easily lead the machine learning model to a skew mapping that only considers high-frequency samples and ignores low-frequency samples. Our PercolationDF breaks through this bottleneck by extracting knowledge into EMKN and medical evidence percolation and has good performance.

In general, our methods outperformed the other three machine learning methods, particularly in $R@10$. Out of the 207 test cases, 51.7% returned all actual diseases in the first 10 results; only 30.0% were returned in a similar previous study [14]. In $P@1$, the other three methods perform similarly and outperform Bayes method. For $R - precision$ and AP, our framework slightly outperforms the others. In general, our framework provides a great improvement in recall and works effectively and steadily.

Moreover, there were 67 records returned with none of the actual diseases in the first 10 results. There may be some reasons for such ineffectiveness: the imperfection of our modified EMKN which consists of some weak causality edges not being filtered, etc. Nevertheless, the shortcoming of multidisease diagnosis is using only symptoms as features; probably, more colorful evidence needs to be applied.

6. Conclusions

The PercolationDF proposed in this paper utilizes an EMKN to effectively model and represent medical knowledge in EMRs and a diagnosis method incorporating percolation theory is increasingly effective in resolving the problem of insufficient data volume where labels have a very strong causality only with respect to certain feature dimensions, and the certain features for different labels are in different dimensions. The clarification of causalities supported in the modified EMKN and the incorporation of percolation theory with our diagnosis method helping not only sufficiently accumulate clinical evidence from graph structure, but also being suitable in inferring on an evolving graph makes a surprising restful progress.

However, the superb result to alleviate insufficient data is frankly based on the limitation of the availability of certain relations from features to labels. The certain relations that other prevailing methods cannot easily obtain are represented in a network effortlessly, but many other kinds of data do not display a certain relation the way an EMR does. Even so, the combination of network and percolation shows great potential for application in the evolution of one system and its different parts.

Future works will focus on making PercolationDF adopt into other situations to model and represent the knowledge that causalities can be realized in the data and infer on the knowledge modeled before, such as some studies on disease complication and deteriorating development between organs and diseases on system of human body.

Acknowledgments

This work is supported by NSFC62006063 and LBH-Z20015.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. M. L. Craig, C. A. Jackel, P. B. Gerrits, Selection of medical students and the maldistribution of the medical workforce in Queensland, Australia, *Aust. J. Rural Health*, **1** (1993), 17–21. https://doi.org/10.1111/j.1440-1584.1993.tb00075.x
2. J. A. Osheroff, J. M. Teich, B. Middleton, E. B Steen, A. Wright, D. E. Detmer, A roadmap for national action on clinical decision support, *J. Am. Med. Inf. Assoc.*, **14** (2007), 141–145. https://doi.org/10.1197/jamia.M2334
3. D. Demner-Fushman, W. W. Chapman, C. J. McDonald, What can natural language processing do for clinical decision support? *J. Biomed. Inf.*, **42** (2009), 760–772. https://doi.org/10.1016/j.jbi.2009.08.007
4. A. N. Kho, J. A. Pacheco, P. L. Peissig, L. Rasmussen, K. M. Newton, N. Weston, et al., Electronic medical records for genetic research: results of the emerge consortium, *Sci. Transl. Med.*, **3** (2011) 79re1. https://doi.org/10.1126/scitranslmed.3001807
5. R. C. Wasserman, Electronic medical recor (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research, *Acad. Pediatr.*, **11** (2011), 280–287. https://doi.org/10.1016/j.acap.2011.02.007
6. A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, *N. Engl. J. Med.*, 2019. https://doi.org/10.1056/NEJMra1814259

7. T. Ma, A. Zhang, AffinityNet: semi-supervised few-shot learning for disease type prediction, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **33** (2019), 1069–1076. https://doi.org/10.1609/aaai.v33i01.33011069
8. Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, preprint, arXiv:1904.05046.
9. M. E. J. Newman, The structure and function of complex networks, *SIAM Rev.*, **45** (2003), 167–256. https://doi.org/10.1137/S003614450342480
10. A. L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: A network-based approach to human disease, *Nat. Rev. Genet.*, **12** (2011), 56–68. https://doi.org/10.1038/nrg2918
11. K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, A. L. Barabási, The human disease network, *Proc. Natl. Acad. Sci.*, 104 (2007), 8685–8690. https://doi.org/10.1073/pnas.0701361104
12. C. A. Hidalgo, N. Blumm, A. L. Barabási, N. A. Christakis, A dynamic network approach for the study of human phenotypes, *PLoS Comput. Biol.*, **5** (2009), e1000353. https://doi.org/10.1371/journal.pcbi.1000353
13. X. Z. Zhou, J. Menche, A. L. Barabási, A. Sharma, Human symptoms–disease network, *Nat. Commun.*, **5** (2014), 4212. https://doi.org/10.1038/ncomms5212
14. C. Zhao, J. Jiang, Z. Xu, Y. Guan, A study of EMR-based medical knowledge network and its applications, *Comput. Methods Programs Biomed.*, **143** (2017), 13–23. https://doi.org/10.1016/j.cmpb.2017.02.016
15. R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, et al., A data mining approach for diagnosis of coronary artery disease, *Comput. Methods Programs Biomed.*, **111** (2013), 52–61. https://doi.org/10.1016/j.cmpb.2013.03.004
16. H. H. Rau, C. Y. Hsu, Y. A. Lin, S. Atique, A. Fuad, L. M. Wei, et al., Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network, *Comput. Methods Programs Biomed.*, **125** (2016), 58–65. https://doi.org/10.1016/j.cmpb.2015.11.009
17. E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, J. Sun, Gram: Graph-based attention model for healthcare representation learning, preprint, arXiv:1611.07012.
18. E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. F. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, (2016), 3512–3520. Available from: https://dl.acm.org/doi/10.5555/3157382.3157490.
19. Z. C. Lipton, D. C. Kale, C. Elkan, R. Wetzell, Learning to diagnose with LSTM recurrent neural networks, preprint, arXiv:1511.03677.
20. F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, J. Gao, 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*, (2017), 1903–1911. https://doi.org/10.1145/3097983.3098088
21. E. Choi, C. Xiao, W. F. Stewart, J. Sun, Mime: Multilevel medical embedding of electronic health records for predictive healthcare, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, (2018), 4547–4557. Available from: https://dl.acm.org/doi/abs/10.5555/3327345.3327366.

22. J. Jiang, X. Li, C. Zhao, Y. Guan, Q. Yu, Learning and inference in knowledge-based probabilistic model for medical diagnosis, *Knowledge-Based Syst.*, **138** (2017), 58–68. https://doi.org/10.1016/j.knosys.2017.09.030
23. D. E. Heckerman, E. J. Horvitz, B. N. Nathwani, Toward normative expert systems: Part I the pathfinder project, *Methods Inf. Med.*, **31** (1991), 90–105. https://doi.org/10.1055/s-0038-1634867
24. J. G. Klann, P. Szolovits, S. M. Downs, G. Schadow, Decision support from local data: creating adaptive order menus from past clinician behavior, *J. Biomed. Inf.*, **48** (2014), 84–93. https://doi.org/10.1016/j.jbi.2013.12.005
25. M. J. Flores, A. E. Nicholson, A. Brunskill, K. B. Korb, S. Mascaro, Incorporating expert knowledge when learning bayesian network structure: a medical case study, *Artif. Intell. Med.*, **53** (2011), 181–204. https://doi.org/10.1016/j.artmed.2011.08.004
26. D. M. Chickering, D. Heckerman, C. Meek, Large-sample learning of bayesian networks is np-hard, *J. Mach. Learn. Res.*, **5** (2004), 1287–1330.
27. T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, preprint, arXiv:1609.02907.
28. C. Y. Wee, C. Liu, A. Lee, J. S. Poh, H. Ji, A. Qiu, et al., Cortical graph neural network for ad and mci diagnosis and transfer learning across populations, *NeuroImage: Clin.*, **23** (2019), 101929. https://doi.org/10.1016/j.nicl.2019.101929
29. R. C. Petersen, P. Aisen, L. A. Beckett, M. Donohue, A. Gamst, D. J. Harvey, et al., Alzheimer's disease neuroimaging initiative (adni): clinical characterization, *Neurology*, **74** (2010), 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25
30. D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, L. Perersson, Graph-based deep learning for medical diagnosis and analysis: past, present and future, *Sensors*, **21** (2021), 4758. https://doi.org/10.3390/s21144758
31. M. Bastian, S. Heymann, M. Jacomy, Gephi: An open source software for exploring and manipulating networks, in *Proceedings of the International AAAI Conference on Web and Social Media*, **3** (2009), 361–362. Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/13937.
32. S. R. Broadbentand J. M. Hammersley, Percolation processes: I. Crystals and mazes, *Math. Proc. Cambridge Philos. Soc.*, **53** (1957), 629–641. https://doi.org/10.1017/S0305004100032680
33. J. M. Hammersley, Percolation processes: II. The connective constant, *Math. Proc. Cambridge Philos. Soc.*, **53** (1957), 642–645. https://doi.org/10.1017/S0305004100032692
34. G. Grimmett, *Percolation*, Springer, New York, 1989. https://doi.org/10.1007/978-1-4757-4208-4
35. E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, J. Sun, Doctor AI: predicting clinical events via recurrent neural networks, in *Proceedings of the 1st machine learning for healthcare conference*, **56** (2016), 301–318. Available from: http://proceedings.mlr.press/v56/Choi16.pdf.
36. 2010 i2b2/va challenge evaluation assertion annotation guidelines. Available from: https://www.i2b2.org/NLP/Relations/assets/Assertion%20Annotation%20Guideline.pdf.
37. 2010 i2b2/va challenge evaluation concept annotation guidelines. Available from: https://www.i2b2.org/NLP/Relations/assets/Concept%20Annotation%20Guideline.pdf.

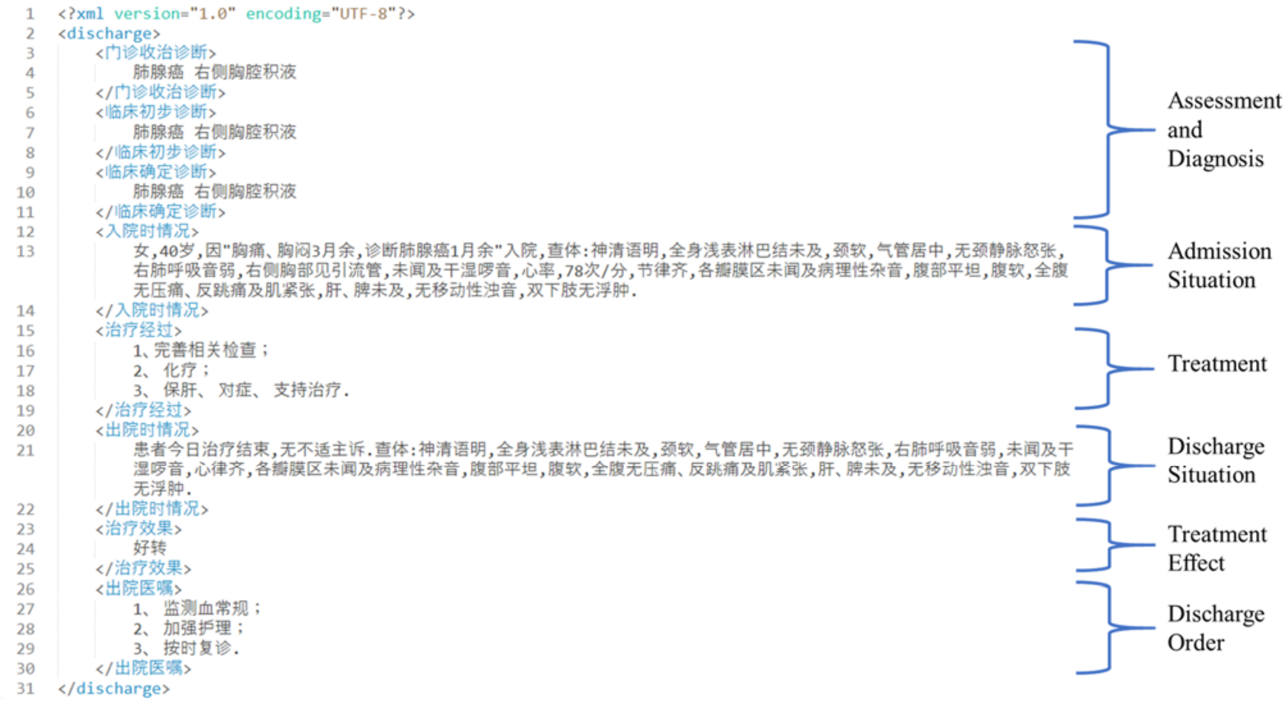
38. J. Yang, Y. Guan, B. He, C. Qu, Q. Yu, Y. Liu, et al., Annotation scheme and corpus construction for named entities and entity relations on Chinese electronic medical records, *J. Software*, **27** (2016), 2725–2746. https://doi.org/10.13328/j.cnki.jos.004880
39. B. He, B. Dong, Y. Guan, J. Yang, Z. Jiang, Q. Yu, et al., Building a comprehensive syntactic and semantic corpus of Chinese clinical texts, *J. Biomed. Inf.*, **69** (2017), 203–217. https://doi.org/10.1016/j.jbi.2017.04.006
40. E. Choi, A. Schuetz, W. F. Stewart, J. Sun, Using recurrent neural network models for early detection of heart failure onset, *J. Am. Med. Inf. Assoc.*, **24** (2017), 361–370. https://doi.org/10.1093/jamia/ocw112
41. P. Nguyen, T. Tran, N. Wickramasinghe, S. Venkatesh, Deepr: a convolutional net for medical records, *IEEE J. Biomed. Health Inf.*, **21** (2017), 22–30. https://doi.org/10.1109/JBHI.2016.2633963
42. C. Zhao, J. Jiang, Y. Guan, X. Guo, B. He, EMR-based medical knowledge representation and inference via Markov random fields and distributed representation learning, *Artif. Intell. Med.*, **87** (2018), 49–59. https://doi.org/10.1016/j.artmed.2018.03.005
43. R. Miotto, L. Li, B. A. Kidd, J. T. Dudley, Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records, *Sci. Rep.*, **6** (2016), 26094. https://doi.org/10.1038/srep26094
44. C. Buckley, E. M. Voorhees, Retrieval evaluation with incomplete information, in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2004), 25–32. https://doi.org/10.1145/1008992.1009000
45. C. Buckley, E. M. Voorhees, Evaluating evaluation measure stability, *ACM SIGIR Forum*, **51** (2017), 235–242. https://doi.org/10.1145/3130348.3130373
46. M. D. Smucker, J. Allan, B. Carterette, A comparison of statistical significance tests for information retrieval evaluation, in *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*, (2007), 623–632. https://doi.org/10.1145/1321440.1321528

Appendix

Table A1 and Figure A1 show a progress note from our records corpus and its corresponding annotation samples.

Table A1. Annotations of the entity and assertion information in the record example.

Medical entity	Position in the text	Entity type
肺腺癌 lung adenocarcinoma	63:66	disease
右侧胸腔积液 dropsy of the right chest	67:73	disease
胸痛 chest pain	172:174	complaintsymptom
胸闷 chest tightness	175:177	complaintsymptom
右肺呼吸音弱 respiratory sounds in right lung is weak	226:232	testresult
引流管 drainage tube	238:241	treatment
干湿啰音 rhonchus and moist rales	245:249	testresult
心率 heart rate	250:252	test
化疗 chemotherapy	354:356	treatment



```
<?xml version="1.0" encoding="UTF-8"?>
<discharge>
    <门诊收治诊断>
        肺腺癌 右侧胸腔积液
    </门诊收治诊断>
    <临床初步诊断>
        肺腺癌 右侧胸腔积液
    </临床初步诊断>
    <临床确定诊断>
        肺腺癌 右侧胸腔积液
    </临床确定诊断>
    <入院时情况>
        女,40岁,因"胸痛、胸闷3月余,诊断肺腺癌1月余"入院,查体:神清语明,全身浅表淋巴结未及,颈软,气管居中,无颈静脉怒张,
        右肺呼吸音弱,右侧胸部见引流管,未闻及干湿啰音,心率,78次/分,节律齐,各瓣膜区未闻及病理性杂音,腹部平坦,腹软,全腹
        无压痛、反跳痛及肌紧张,肝、脾未及,无移动性浊音,双下肢无浮肿.
    </入院时情况>
    <治疗经过>
        1、完善相关检查；
        2、 化疗；
        3、 保肝、 对症、 支持治疗.
    </治疗经过>
    <出院时情况>
        患者今日治疗结束,无不适主诉.查体:神清语明,全身浅表淋巴结未及,颈软,气管居中,无颈静脉怒张,右肺呼吸音弱,未闻及干
        湿啰音,心律齐,各瓣膜区未闻及病理性杂音,腹部平坦,腹软,全腹无压痛、反跳痛及肌紧张,肝、脾未及,无移动性浊音,双下肢
        无浮肿.
    </出院时情况>
    <治疗效果>
        好转
    </治疗效果>
    <出院医嘱>
        1、 监测血常规；
        2、 加强护理；
        3、 按时复诊.
    </出院医嘱>
</discharge>
```

Figure A1. A discharge note from the records corpus.