Similarity Computing on Electronic Health Records

Completed Research Paper

Suresh Pokharel
The University of Queensland
Australia
s.pokharel@uq.edu.au

Xue Li
The University of Queensland
Australia
xueli@itee.uq.edu.au

Xin Zhao
The University of Queensland
Australia
uqxzhao1@uq.edu.au

Anoj Adhikari
Dhaulagiri Zonal Hospital
Nepal
anojadhikari@sjtu.edu.cn

Yu Li
The University of Queensland
Australia
yuli@itee.uq.edu.au

Abstract

Similarity computing on real world applications like Electronic Health Records (EHRs) can reveal numerous interesting knowledge. Similarity measures the closeness between comparable things such as patients. Like similarity computing amongst Intensive Care Unit (ICU) patients can create various benefits, such as case based patient retrieval, unearthing of similar patient groups. However, many classical methods such as euclidean distance, cosine similarity can't be directly applicable as similarity computing in EHRs is subjective and in many cases conditional. Also, many intrinsic relationships between the data are lost due to poor data representation and conversion. To address these challenges, firstly, we propose structural network representation for EHRs to preserve inherent relationship. And, to make them more comparable, we do data enrichment e.g. adding abstract information. Then, we extract different similarity feature sets to generate different similarity metrics and retrieve top similar patients. Finally, we perform experiment which shows promising results over classical methods.

Keywords: EHR, Multivariate, Patient Similarity, Similarity Computing

Introduction

Large amount of data are collected by many organizations such as hospital, bank, government, industry, business organization etc. for future use. This dramatic growth of data creates many opportunities as well as challenges. Similarity computing is one of the fundamental problem for knowledge discovery. Similarity measures the closeness, or "sameness", of two comparable objects (Mo et al. 2013). If it is addressed properly, many real world problem can be solved. For example, in *EHRs*, finding similar previous cases can play vital roles in many aspects of patient management such as: if we can provide information of previous patients who have had similar signs and symptoms as the new patient, then the treating physician or the team will have more information regarding the disease process leading to early detection of disease pattern in the new patient. This will help in narrowing the provisional diagnosis, planning the correct course of treatment, selecting the necessary routine and special investigations and

prescribing correct dose of medicine. This will also support the treating team to remain more vigilant and prepared for any disease complication or detection of new symptoms at the earliest. All these can lead them to take quick and confident decisions with better outcome. Finding patient cohort is useful for comparative effectiveness studies and better understanding of patients (Sun et al. 2012).

Many traditional similarity computing methods such as cosine similarity, euclidean distance, Heterogeneous Euclidean Overlap Metric function (HEOM) (Wilson et al. 1997) are well suited for simple data but don't consider the complex data which can be structured in a network. The semantic relation between the concepts can be realized in the structural network representation. Network similarity methods such as the co-citation (Small 1973) calculate the similarity between web pages A and B by number of pages cite both page A and page B whereas bibliographic coupling (Kessler 1963) measures the similarity between web pages A and B by number of pages cited by both page A and B. Kleinberg (1999) presented the authority and hub matrices for citation analysis and connection with co-citation and co-reference. However, the EHRs data possess more complicated structure and traditional methods are not adequate for representation as well as similarity computation.

In this research, a new approach for similarity computing in EHRs is presented in the context of Intensive Care Unit (ICU) patients. More specifically, we aim to solve the following problem; "*Given the patients in ICU, how to retrieve the top k similar patients for a given query patient?*" In order to achieve this, we proceed with the following steps.

- ***Representation of EHR into structure network:*** Inherent data relations are present in ICU data. Structural network representation is one of the best way to preserve those relations. Resource Description Framework (RDF) technology[1] is used for structural network representation (see methodology for more details) due to the following advantages: (1) It is simple yet flexible schema model where any piece of information can be easily inserted and joined with exiting data. (2) This dataset can be easily linked with other heterogeneously distributed data set via linked open data (LOD)[2] for richer information.
- ***Data enrichment:*** To make data more comparable for similarity computing, we add additional information such as summary information by applying semi-automatic data enrichment process (see methodology for more details).
- ***Feature extraction:*** Many features such as different biochemical measurements, treatments etc. are contained in ICU data which are useful for similarity computing. To study the results of each feature set and combine feature sets in terms of similarity, we extract nine different feature sets and two combined feature sets (see methodology for more details).
- ***Similarity computing:*** Similarity matric is generated based on the important features shared by patients in order to find the top *k* similar patients for a given query patient (see methodology for more details).

Research Contribution: The contributions of this research are as follows:

1. We show the structural network representation of EHR called Structure ICU Semantic Network (*SISNet*).
2. We present different feature sets which contribute similarity computing.
3. We propose a new method called *Inherent Similarity* (*ISim*) to produce patient similarity matrix.
4. We show the good results which will support health professionals.

The remainder of this paper is structured as follows: in the subsequent section, we present the related works followed by a detailed description of the methodology. Then, in experiment section, we describe the data sources being used, perform experiment, evaluate the work, and give concluding remarks.

[1] https://www.w3.org/RDF/

[2] http://linkeddata.org/

Related Work

We organize the related works for computing similarity into two sections: firstly, we will discuss about patient similarity techniques and secondly, semantic similarity techniques.

Patient Similarity: Sun et al. (2012) presented patient similarity matrix by using Locally Supervised Metric Learning (*LSML*); secondly, they showed the update mechanism to existing matrix by relating eigenvalue and eigenvector and finally, they proposed Composite Distance Integration (Comdi) for integrating multiple physician's feedback metrics to similarity matrix. Chan et al. (2010) extracted 14 different similarity measures of patient for classifying the HCC patient and use support vector machine for supervised learning and named as *SimSVM*. This type of supervised learning is good only for labeled data. It requires data conversation and needs to handle missing data effectively and also ignored many attributes. And it is not designed for similarity computing for structured networks. Jiang et al. (2014) presented treatment data into linked open data and linked with Linked Life Data (LLD)[3] to integrate the share data. The patient similarity is calculated based on the semantic distances of drugs taken by patient and drug similarity is the semantic distance between the doctor's advice and terminology in LLD. In this research, they only consider treatments and don't consider complex relationships.

Semantic Similarity: Basically, there are two approaches for measuring semantic similarity (Zhu et al. 2017; Sanchez et al. 2012); (1) Corpus-based approaches: The similarity between two words is measured in terms of information gain for their surrounding words. That means, two words are more similar if there is more overlapping between surrounding words. It is simple and easy to implement. But, it doesn't take account the semantic meaning and structure of the graphs. (2)Knowledge-based approaches: Here, the shortest path between the concepts is calculated in the graph based on their location. That means, lower level concepts are more similar than higher level concepts. It has ability to capture the semantic similarity. Mheich (2017) considered physical location of node and proposed a similarity method called *SimiNet* for quantifying brain network similarity. However, ICU data are more complex due to multivariate and temporal nature. So, still need further improvement in existing state of art techniques.

Methodology

The main concept behind the similarity between two patients is "*if they share more number of important features like treatments, laboratory tests then they are more similar*". In this section, firstly, we explain structural network representation, then data enrichment process followed by features extraction and finally, describe similarity computing method.

Representation of EHRs into Structural Network

Intrinsic associations between the data are exist in EHRs. Preserving these relations helps to better understand the data. One of the best way is to represent them into structural network. Originally, the available ICU data is present in relational database. We propose RDF technology for data representation. After the conversion, we get the structural network representation of ICU data and called it as **S**tructure **I**CU **S**emantic **Net**work (***SISNet*)**. The procedure for conversion is as follows:

Requirements

The following requirements are satisfied during the conversion:

- *1:1 mapping:* Each data value of relational database is represented into RDF triple so that there is no chance of missing information.
- *All the constraints most satisfied:* Relational database constraints such as primary keys, foreign keys is preserved as domain and range in RDF so that there is no chance of losing semantics of database.

[3] http://linkedlifedata.com/

The purpose of the above requirements is to preserve all the information so that during reverse conversion, we can get the same dataset.

Generating Internationalized Resource Identifiers (IRIs)

IRIs are used for representing each piece of data uniquely and we follow the direct mapping[4] (Arenas et al. 2012) to convert Relational Database (RDB) to Resource Description Framework (RDF). IRIs are generated as follows:

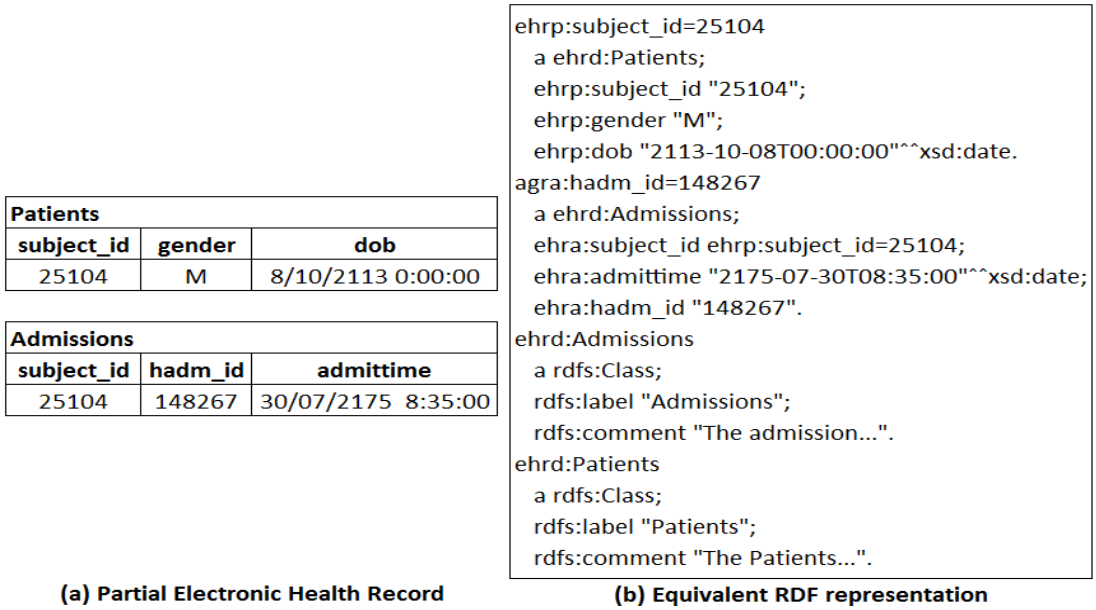
- IRIs for tables are generated by concatenating base *IRI*, *vocab* word with *table name*. The *base_IRI* is part of IRI which is common to all. The format is as follows: *base_IRI/vocab/table_name.*
- IRIs for table column are generated by concatenating *base IRI*, *vocab world*, *table name* and *column name*. The format is as follows: *base_IRI/vocab/table_name#column_name*
- IRIs for each row are generated by concatenating *base IRI*, *data word*, *table name*, *column name* of primary key and value of row in that column. The format is as follows: *base_IRI/data/table_name/column_name=value*
- Other vocabularies: Other vocabularies are generated as format *base_IRI/vocab/other_vocabularies.*

Mapping Rules

For mapping each data into RDF, we satisfy the following mapping requirements:

- Every row of the table generates a set of RDF triples.
- Each table name is taken as class name.
- Each column (attribute) of table is a property which has its domain and range.

Figure 1 shows an example of converting relational database into RDF. Here, *Patients* and *Admissions* are tables where *subject_id* and *hadm_id* are primary keys respectively.



Patients

subject_id	gender	dob
25104	M	8/10/2113 0:00:00

Admissions

subject_id	hadm_id	admittime
25104	148267	30/07/2175 8:35:00

```
ehrp:subject_id=25104
  a ehrd:Patients;
  ehrp:subject_id "25104";
  ehrp:gender "M";
  ehrp:dob "2113-10-08T00:00:00"^^xsd:date.
agra:hadm_id=148267
  a ehrd:Admissions;
  ehra:subject_id ehrp:subject_id=25104;
  ehra:admittime "2175-07-30T08:35:00"^^xsd:date;
  ehra:hadm_id "148267".
ehrd:Admissions
  a rdfs:Class;
  rdfs:label "Admissions";
  rdfs:comment "The admission...".
ehrd:Patients
  a rdfs:Class;
  rdfs:label "Patients";
  rdfs:comment "The Patients...".
```


Figure 1. Partial Record of rdb2rdf Conversion

[4] https://www.w3.org/TR/rdb-direct-mapping/

Data Enrichment

Making data comparable is one of the challenging task in EHR and enriching data will help for this. Two kinds of meta-information is added by applying semi-automatic process.

- ***Value abstraction:*** Many attributes such as laboratory results, sensor reading (e.g. body temperature, blood pressure etc.) are expressed in continuous values. They indicate the patient's situation and ideally be in normal range. However, these indicators might deviate from the normal range. It is useful to know the degree of deviation in different ranges. For example, stages of chronic kidney disease (CKD) can be graded according to *glomerular filtration rate* (GFR) as Stage 1: >= 90ml/hr, Stage 2: 89-60ml/hr, Stage 3: 59-30ml/hr, Stage 4: 29-15ml/hr and Stage 5: < 15ml/hr. If two patients have their GFR less than 15 ml/hr, then they may be similar and kidney failure patients. While in other patients with uncontrolled diabetes mellitus, it may lead to osmotic diuresis and hence significantly increase GFR leading to polyuria. Likewise, diabetes insipidus also causes abnormally high GFR. In order to make them comparable, we add five interval scales for continuous value representation for laboratory test outcome which are ; *Very Low (VL), Low (L), Normal(N), High(H), Very High(VH)*. A statistical method, percentile method (Batal et al. 2013) is applied due to the unavailability of value range information in database. The interval scales are defined from the global data as follows: below 10^{th} percentile is *VL*, 10^{th} to 25^{th} percentile is *L*, 25^{th} to 75^{th} percentile is *N*, 75^{th} to 90^{th} percentile is *H* and above 90^{th} percentile is *VH*.
- ***Abnormality information:*** Abnormality information are very useful for diagnosing diseases and many other purposes, like blood sugar level to diagnose diabetes, serum creatinine level to diagnose kidney disease, total white blood cell (WBC) count for patient's infection status, etc. To capture these kinds of indicators, the normal and abnormal values of laboratory test are marked as *normal* and *abnormal* respectively. The abnormal information can be obtained from database. If database doesn't contain abnormal value, we assume it as normal value.

Feature Extraction

Different features are extracted for similarity computing. In this study, we aim to investigate different features sets and their performances. Hence, we extracted the following different feature sets from *SISNet*.

- ***Chart events:*** Chart events contain all the routine vital signs and other related information like mental status, ventilator settings etc. It covers bulk portion of information.
- ***Qualitative:*** The laboratory test results contain both qualitative data as well as quantitative values. The laboratory outcome such as positive, negative etc. are taken as qualitative feature set.
- ***Abstract:*** Abstract values (*VH, H, N, L,* and *VL*) are generated as described in previous subsection and taken as abstract feature set.
- ***Abnormality:*** The laboratory test outcomes may be normal or abnormal. Abnormality feature set is generated as described in previous subsection.
- ***Prescription:*** The medicines taken by patients are captured under prescription feature set.
- ***Output events:*** The output events feature set contains the output such as urine or fluid extracted from drain of the patients.
- ***Microbiology events:*** Patient's microbiology tests during hospital stay are taken as microbiology feature set.
- ***Laboratory events:*** Patient's laboratory measurements during hospital stay are considered as laboratory feature set.
- ***Input events:*** This feature set carries the information about any fluid which have been given to patients.
- ***9Featureset:*** 9Featureset is generated by considering all of nine feature sets as described above. It includes chart events, qualitative, abstract, abnormality, prescription, output events, microbiology events, laboratory events and input events. The motivation of this feature set is to know the result of combine feature sets.

- ***7Featureset:*** 7Featureset is generated by considering seven feature sets as describe above. It includes chart events, qualitative, abstract, abnormality, output events, microbiology events, laboratory events. This feature set doesn't include treatment related features such as prescriptions and input events. This type of feature without treatment will be useful for treatment recommendation based on different observations.

The feature sets contain two kinds of information (1) value related information and they are *qualitative, abstract, abnormality* (2) item related information which consider only frequency of item appearance in patients and doesn't consider amount, values information and they are *chart events, prescriptions, output events, microbiology events, laboratory events, input events*. The combine feature sets *9Featureset* and *7Featureset* are useful for analyzing combine effect of different feature sets. In nutshell, the similarity of patients depends on the number of common important features from the above similarity feature sets.

Similarity Computation

One of the main contribution of this research is similarity computing. From the different feature sets as explained in above section, we generate the different *patient similarity metrics* by using the method below. Here, first, we will discuss about similarity computing method, give explanation and finally describe with an example.

Patient Similarity matrix

The ICU patients contain the Inherent Similarity, denoted as *ISim*, and the similarity between two patients *a* and *b* is denoted as $ISim(a,b)$ and calculated as the sum of total number of common features with their frequencies share by both. We adopt the idea of *tf*idf* method (Salton and Buckley, 1988) and modified in our case and express in Equation 1.

$$ISim(a,b) = \frac{\sum_{i=1}^{n} \min(\left(f_{a_i} * w_i\right), \left(f_{b_i} * w_i\right))}{avg(\sum_{i=1}^{n}\left(f_{a_i} * w_i\right), \sum_{i=1}^{n}\left(f_{b_i} * w_i\right),)} \quad (1)$$

Where, $f_{a_i}$and $f_{b_i}$are the i^{th} feature frequencies of patient *a* and *b*; n is the total number of features; w_i is the overall importance of i^{th} feature and defined in Equation 2. Here, we take minimum (min function) number of frequencies of common features shared by both patients. That is if one patient has taken drug (say diazepam) 15 times and another patient take same drug 7 times, we take minimum which is 7. The average (avg function) is taken for normalization. The range of $ISim(a,b)$ is *0* to *1*.

$$w_i = log_{10} \frac{(\mathrm{M}+1)}{(m_i+1)} \quad (2)$$

Where, *M* is the total number of patients and m_i is number of patients containing i^{th} feature. log_{10} is taken for normalizing value to smaller value. One is added to avoid the possible values of *0* or *undefined*.

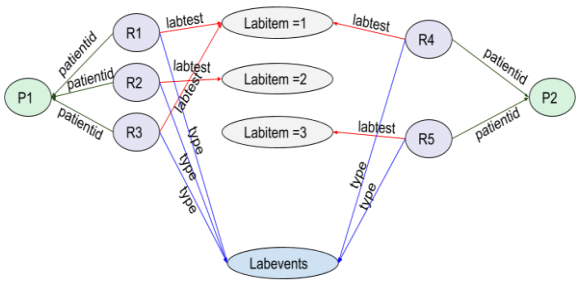
Explanation

The similarity method is based on two things; (1) how many features are shared (2) what is the overall importance of feature? We give less weight to the features which are more common and give more weight to those features which are uncommon. For example, fever can be the foremost symptom for many diseases and may occur in a large number of patients with different illnesses. Likewise, the overall importance of prescription like *normal saline, dextrose solution, Acetaminophen* etc. is low as these are generally given to every ICU patients. While prescriptions like epinephrine, diazepam, insulin, adenosine will have higher overall importance as they are used in special situations and for a particular group of patients in ICU. In general, the patients are considered to be more similar if they share more number of uncommon features in higher frequencies. For example, there is very high chance that two feverish patients with daily prescription of *hypoglycemic* drugs and recent urine culture, total blood

count and fasting blood sugar test are most probably suffering from urinary tract infection. But fever alone can be caused by different diseases like *enteric fever, malaria, encephalitis, upper respiratory tract infection* etc., and *total blood count* can be requested in all these diseases.

An Example

Figure 2 shows an example for calculating the similarity between two patient *P1* and *P2*. *R1, R2, R3, R4 and R5* are the record numbers, *Labitem=1, Labitem=2, and Labitem=3* are the laboratory test items (such as cholesterol, globulin etc.) of *Labevents*. Figure 2(b) shows the corresponding frequencies of laboratory test done by patient *P1* and *P2*. Figure 2(c) illustrates the calculation of overall importance of labitem features by using equation 2. By using equation 1, the similarity between *P1* and *P2* is *0.3556*.



(a) An Example for Labevents Test Done by Two Patients

Patient\Labevents	labitem=1	labitem=2	labitem=3
P1	2	1	0
P2	1	0	1

(b) Patients and Features Matrix

Features	labitem=1	labitem=2	labitem=3
weight	0.564271	0.7403627	0.7403627

(c) Importance of Features Calculation (let M=10)

Figure 2. An Example for Patient Similarity Calculation

Experiment

Experiment Setup

Dataset:

An open patient's medical data, MIMIC-III[5] is used for this research. It contains real time sensor data as well as history data; it contains 53,423 distinct adult patient (aged 16 years or above) hospitalized in the ICU between 2001 and 2012 (Johnson et al. 2016). It has 46520 patients where 26121 are male and 20399 are female. After converting *MIMIC-III* dataset which is available in relational database, the resulting *SISNet* contains more than 5.06 billion triples. *Table 1* shows the number of triples as well as distinct subjects falling into different important classes.

Tools

API of Apache Jena[6] is used for relational database conversion. Apache Jena Openlink Virtuoso[7] is used for storing and querying RDF.

Patient Selection

We empirically select patient admissions from four different groups based on gender and age (World Health Organization 1982) as follows; (1) Group I: *older adulthood to average retirement [45 to 65] male patients* (2) Group II: *older adulthood to average retirement [45 to 65] female patients* (3)

[5] https://mimic.physionet.org/

6 https://jena.apache.org/
7 https://virtuoso.openlinksw.com/

Group III: *retirement [65 to 90] male patients* (4) Group IV: *retirement [65 to 90] female patients*. The reason behind this selection is both age and sex affect body physiology and hence have different effects on

Table 1. SISNet Triples Details

Data	# Triples	# Distinct Subjects
Admissions	1,164,665	58,976
Chartevents	4,004,117,513	330,712,483
Microbiologyevents	9,196,590	631,726
Inputevents_cv	330,632,138	17,527,935
Inputevents_mv	112,430,034	3,618,991
Outputevents	60,596,693	4,349,218
Prescriptions	88527820	4,156,450
Labevents	336,941,930	27,854,055
Items	99,722	12,487
Icustays	926,696	61,532
Others	116,235,014	8,879,069
Total	**5,060,868,815**	**397,862,922**

behavior and disease. For example, uterine prolapse only occurs in female patient while prostate cancer occurs only in male patients and generally in old age. Here, each group contains top 10 different common diagnoses. There is possibility of having more than one hospital admission for a patient and they may be diagnosed with different diseases. For example, we observed that a male patient had been admitted to the hospital six times during his life time and diagnosed one time with *FEVER; TELEMETRY*, three times with *SEPSIS*, one time with *CEREBELLAR MASS* and once with *PNEUMONIA; HYPOXIA*. So, we treat each hospital admission as distinct patients and we have total 7416 distinct patients consisting of 1586, 1211, 2405, 2214 patients in each Groups I-IV respectively.

Evaluation Plan

International Classification of Diseases, Ninth Revision (icd9)[8] diagnosis is standardized and accepted worldwide. So, we take patient's disease diagnosis as the ground truth to evaluate our method. Each diagnosis has unique icd9 code. Similarity for our ground truth is "*two patients are similar if they diagnosed with the same disease*". One patient can have multiple diagnoses at each hospital admission. In data, we observe that there is 6984 distinct icd9 code and maximum 39 diagnosis for a patient in one hospital admission. There are sequence numbers (here, we sometimes referred as order and increasing sequence number means decreasing order) associated with icd9 code for every patient which signifies the priority of diagnoses[9] (Health Statistics 2011). In other words, higher order diagnosis is the primary reason for patients being admitted to the hospital. In order to make multiple diagnoses comparable, we assign the weight for each diagnosis based on priority by applying modified sigmoid function. Sigmoid function is chosen to reduce significance of diagnosis with decreasing diagnosis order as shown in equation 3.

$$w_i = \frac{2}{(e^{(N-1)}+1)} \tag{3}$$

Where, w_i is the weight of i^{th} diagnose having sequence number N.

The similarity between two patients a and b is called as similarity Index (*simIndex*) and calculated as shown in equation 4.

8 https://www.cdc.gov/nchs/icd/icd9cm.htm
9 https://mimic.physionet.org/mimictables/diagnoses_icd/

$$SimIndex(a,b) = \frac{\sum_{i=1}^{n} \min(a_{w_i}, b_{w_i})}{avg(\sum_{i=1}^{n} a_{w_i}, \sum_{i=1}^{n} b_{w_i})} \quad (4)$$

Where, a_{w_i} and b_{w_i} are the weights of i^{th} diagnoses for patient a and b respectively and calculated using equation 3; n is total number of diagnoses for patients a and b. We take minimum value (min function) to capture similarity if they share the common diagnoses and average (avg function) for normalization. The range for *SimIndex* is *0* to *1*.

Example 1**:** *Table 2 shows the top 5 diagnoses of two patient a and b and associated sequence number is shown in corresponding seq_number column. Then, the corresponding weights a_w and b_w are calculated by using Equation 3 and simIndex between two patients is 0.193606797 which is calculated by using Equation 4. In this case, three diagnoses of patient b having sequence number 1, 4 and 2 is matched with three diagnoses of patient a having sequence number 3, 4 and 5 respectively and produce simIndex 0.193606797. So, the higher the simIndex more the similarity between patients. simIndex greater than 0 means there is at least some common diagnoses otherwise no overlapping.*

Table 2. Example for Similarity Index Calculation

ICD9 Code	Patient a		Patient b		
	Seq_number	ICD9 weights: a_w	Seq_number	ICD9 weights:b_w	Min(a_w,b_w)
5789	1	1	-	0	0
389	2	0.537882843	-	0	0
5070	3	0.238405844	1	1	0.238405844
51881	4	0.094851746	4	0.094851746	0.094851746
78552	5	0.03597242	2	0.537882843	0.03597242
99592	-	0	3	0.238405844	0
7070	-	0	5	0.03597242	0
Sum		**1.907112853**	-	**1.907112853**	**0.36923001**
				SimIndex(a,b)	0.193606797

A disease is a very complex physiology affecting different systems and organs of our body to different extent. Our body systems are so inter related that the problem in one system may cause symptom in an organ more related to another system, e.g. problems with kidneys may lead to heart disease. Somebody diagnosed with diabetes maybe taking hypoglycemic drug for better regulation of his daily blood sugar level, but at the same time is more prone to heart disease or kidney disease or eye problems. Likewise, somebody diagnosed with hypertension and regularly taking anti-hypertensive drug is also more prone to disease of the heart, kidney or eyes. Even though their primary diagnosis and primary prescription are different, these two patients often end up in the hospital with similar problems and hence may have many similar lower order diagnosis.

Experiment Procedure

Listing 1 shows the experiment steps. The features are extracted and then the patient similarity matrix is constructed as described in methodology. The Similarity@k for each patient is calculated in two steps: (1) retrieving top k similar patients (2) finding similarity indexes with all k patients and taking average of them. Then, for plotting the *Similarity@k* versus *Patient (%)* graph: we divide similarity range (0 to 1) into different m (in our case, m=1000) values, to show the result by varying the similarity@k, and find the number of patients (%) having similarity@k greater than or equal to m value.

Results and Discussions

To compare the performance of different feature sets, we plot Similarity@k versus patient (%) graph (as describe in previous subsection) for top three (k=3) similar patients which is shown in *Figure 3*. Due to high number of feature sets, we divided them into two sets to avoid overlapping so that each line is clearly visible.

```
begin:
Extract different features
Construct Patient Similarity Matrix          \\using equation 1
for each patient P, i: 1 to n                \\n is number of patients
   Pi_list ← {}                              \\Initialize empty similarity Index list
   retrieve top k similar patient
   for j: 1 to k
      P_ij ← simIndex(Pi, Pj)                \\using equation 4
      add P_ij to Pi_list
   Similarity@k(Pi) ← average(Pi_list)
for each data point m:                       \\ Different m ( let m=1000) values in 0 to 1
   Calculate patients (%) having Similarity@k (Pi) ≥ each data point.
   Plot: Similarity@k Vs patients (%)
end
```

Listing 1. Experiment Steps

Certainly, two patients with same diagnosis are more similar, but at the same time, prescription for a disease for same age group and sex of patient is also very alike. So, we can take the ICD9Similarity as ground truth. The graph portrays that the *9Features, 7Features, inputevents, abstract features* are performing well slightly below than *prescription*. The graph also depicts the patient (%) is linearly decreasing with increasing *simIndex*. Similarly, nearly 80 percent of patients meet *simIndex* 0.19 which means there are at least some overlapping with few higher order diagnoses.

We compare our method (*ISim*) with other two widely used traditional similarity learning methods: *tf*idf with cosine* and *tf*idf with euclidean* which is shown in *Figure 4*. We select two top performing features (*prescription* and *inputevent*) for comparison. The graph illustrates that *cosine* method is performing better then *euclidean* method and our method *ISim* is better than both classical methods.

Here what we need to understand is that in real patient population, the number of patients with same first diagnosis is generally less than the number of patients with overlapping diagnosis without considering order. Most of the time the patients generally have common lower order diagnosis. Our aim is to identify the most number of patients with maximum number of rare similar features. So even with *simIndex* of about *0.19*, we can identify a pool of patients with multiple similarities amongst them. Our focus is the diagnosis and treatment of present illness. So, from these patients, if we can identify couple of patients with similarities most matching to the present situation of our patient, it would be a great achievement and a large support for health professionals; which is clearly shown by the graphs.

Conclusion and Future Work

In this paper, we presented the similarity computing on EHR in the context of ICU patient similarity. To achieve this, firstly, we presented *SISNet* for representation of ICU data into structure network. Secondly, we presented a method called *ISim* for generating different similarity metrics by computing similarity from different extracted features sets. Finally, we presented an objective evaluation method to verify our method. We obtained promising results which will be helpful for health professionals.

For future work, many temporal relations such as before, after, during etc are exit in ICU data. These relations can be preserved and represented (Batsakis et al. 2017) in structure network. We plan to incorporate the techniques to capture those patterns. Furthermore, ICU patients follow sequences like observation, treatment, observation, treatment during hospital stay. During this process, there is complex relationship present between multiple observations and treatments with temporal relation. So, we plan to extend our work to investigate such patterns.

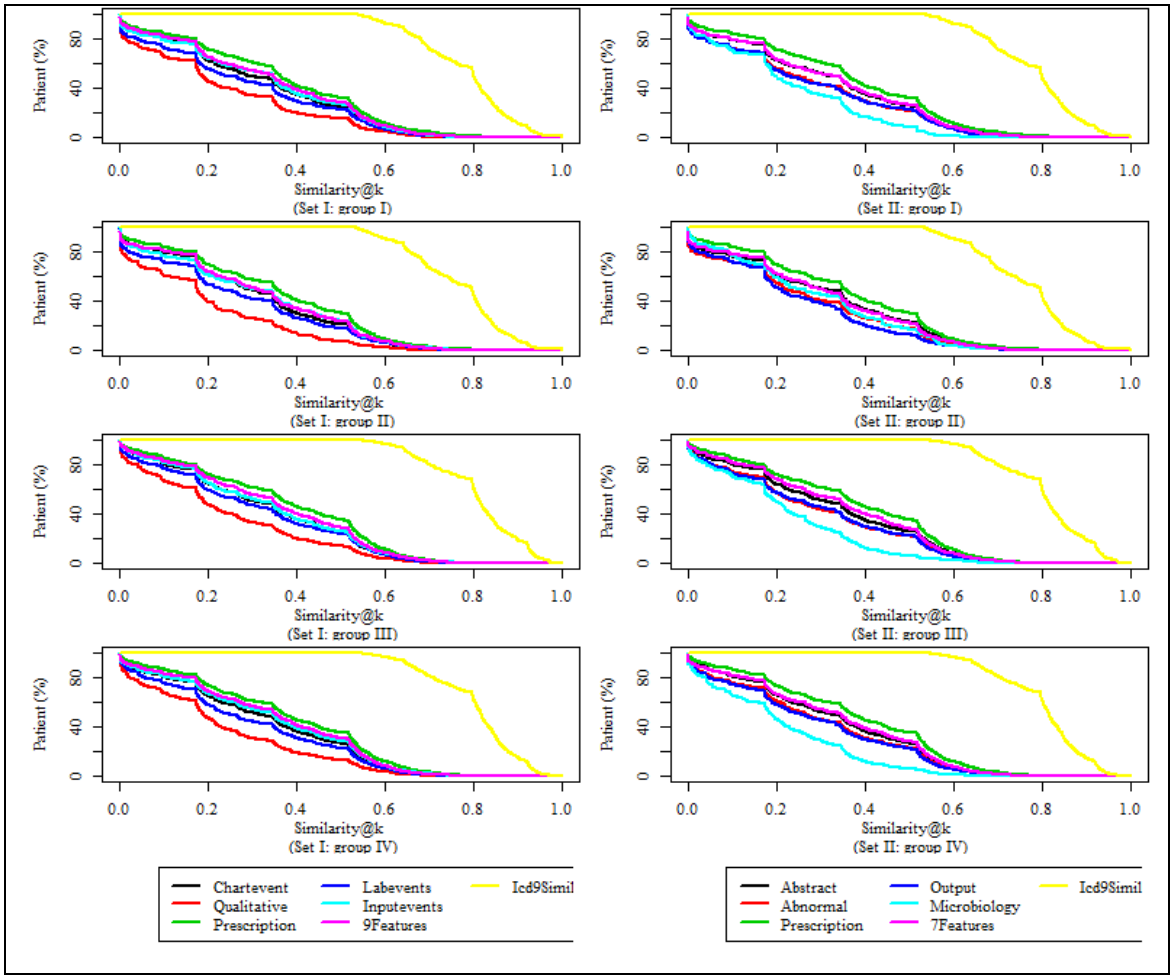


Figure 3. Patient Similarities Comparison for Different Feature Sets at k=3

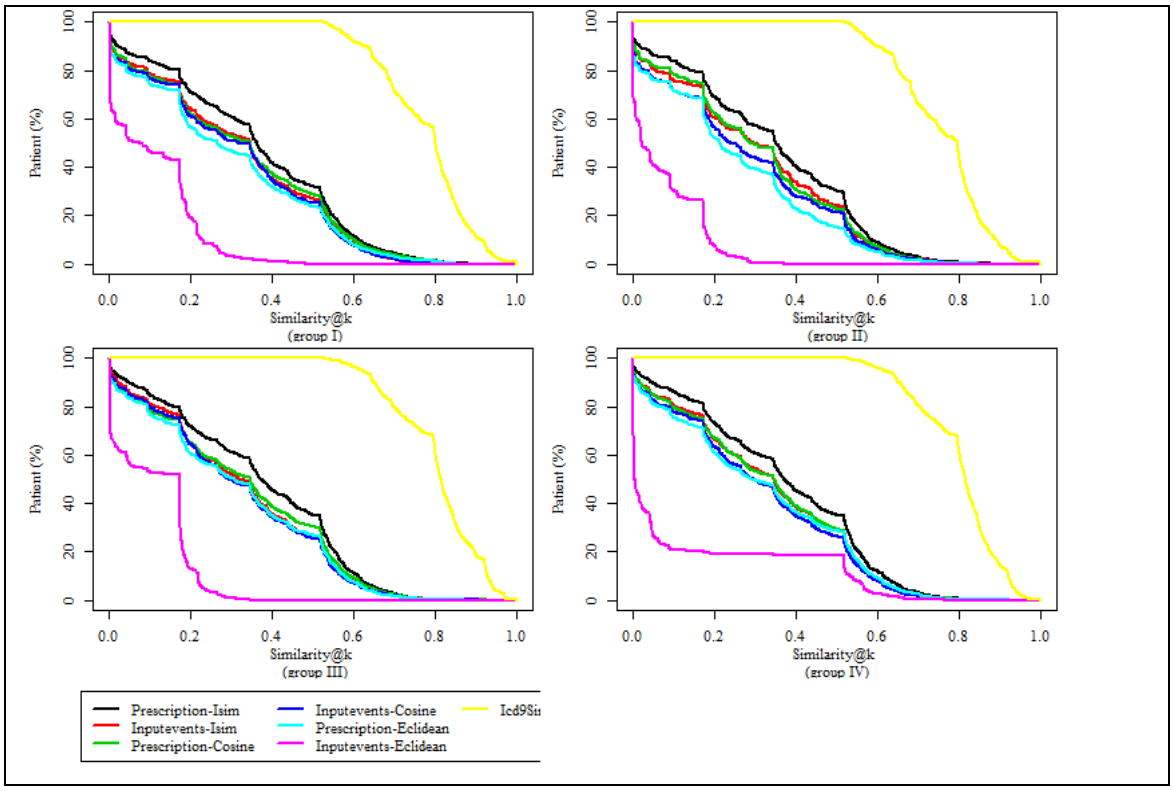


Figure 4. Comparison of Isim with Other Methods

References

Arenas, M., Bertails, A., Prud, E. and Sequeda, J. 2012. "A direct mapping of relational data to RDF," *W3C*.

Batal, I., Valizadegan, H., Cooper, G.F., and Hauskrecht, M. 2013. "A temporal pattern mining approach for classifying electronic health record data," in *ACM Transactions on Intelligent Systems and Technology* (4:4), p. 63.

Batsakis, S., Petrakis, E.G., Tachmazidis, I., and Antoniou, G. 2017. "Temporal representation and reasoning in OWL 2," *Semantic Web* (8:6), pp. 981-1000.

Chan, L.W.C., Chan, T., Cheng, L.F., and Mak, W.S. 2010. "Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy," in *IEEE International Conference on Bioinformatics and Biomedicine Workshops*, pp. 467-470.

Jiang, L., Li, L., Cai, H., Liu, H., Hu, J., and Xie, C. 2014. "A linked data-based approach for clinical treatment selecting support," *Journal of Management Analytics* (1:4), pp. 301-316.

Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., and Mark, R.G. 2016. "MIMIC-III, a freely accessible critical care database," *Scientific data* (3), p. 160035.

Kessler, M.M. 1963. "Bibliographic coupling between scientific papers," *Journal of the Association for Information Science and Technology* (14:1), pp. 10-25.

Kleinberg, J.M. 1999. "Authoritative sources in a hyperlinked environment," *Journal of the ACM* (46:5), pp. 604-632.

Mheich, A., Hassan, M., Khalil, M., Gripon, V., Dufor, O., and Wendling, F. 2017. "SimiNet: a Novel Method for Quantifying Brain Network Similarity," in *IEEE transactions on pattern analysis and machine intelligence*.

Mo, R., Ye, C., and Whitfield, P. H. 2013 "Some Similarity Indices with Potential Meteorological Applications," *National Laboratory for Coastal and Mountain Meteorology*, Technical Report, No. 2013-002.

National Center for Health Statistics and Centers for Medicare and Medicaid Services. 2011. "ICD-9-CM official guidelines for coding and reporting," Washington (DC): US GPO.

Salton, G., and Buckley, C. 1987. "Term weighting approaches in automatic text retrieval," *Cornell University*.

Sánchez, D., Batet, M., Isern, D., and Valls, A. 2012. "Ontology-based semantic similarity: A new feature-based approach," *Expert Systems with Applications* (39:9), pp. 7718-7728.

Small, H. 1973. "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the Association for Information Science and Technology* (24:4), pp. 265-269.

Sun, J., Wang, F., Hu, J., and Edabollahi, S. 2012. "Supervised patient similarity measure of heterogeneous patient records," in *ACM SIGKDD Explorations Newsletter* (14:1), pp. 16-24.

Wilson, D.R., and Martinez, T.R. 1997. "Improved heterogeneous distance functions," *Journal of artificial intelligence research* (6), pp. 1-34.

World Health Organization, 1982. "Provisional guidelines on standard international age classification," *Department of International Economic and Social Affairs*. Series M, (74).

Zhu, G., and Iglesias, C.A. 2017. "Computing semantic similarity of concepts in knowledge graphs," in *IEEE Transactions on Knowledge and Data Engineering* (29:1), pp. 72-85.