Mining Diabetes Complication and Treatment Patterns for Clinical Decision Support

Lu Liu[‡], Jie Tang[†], Yu Cheng[‡], Ankit Agrawal[‡], Wei-keng Liao[‡], Alok Choudhary[‡]
[‡]EECS Department, Northwestern University
[†]Department of Computer Science and Technology, Tsinghua University
[‡]{llg183, ych133,ankitag,wkliao,choudhar}@eecs.northwestern.edu
[†]jietang@tsinghua.edu.cn

ABSTRACT

The fast development of hospital information systems (HIS) produces a large volume of electronic medical records, which provides a comprehensive source for exploratory analysis and statistics to support clinical decision-making. In this paper, we investigate how to utilize the heterogeneous medical records to aid the clinical treatments of diabetes mellitus. Diabetes mellitus, simply diabetes, is a group of metabolic diseases, which is often accompanied with many complications. We propose a Symptom-Diagnosis-Treatment model to mine the diabetes complication patterns and to unveil the latent association mechanism between treatments and symptoms from large volume of electronic medical records. Furthermore, we study the demographic statistics of patient population w.r.t. complication patterns in real data and observe several interesting phenomena. The discovered complication and treatment patterns can help physicians better understand their specialty and learn previous experiences. Our experiments on a collection of one-year diabetes clinical records from a famous geriatric hospital demonstrate the effectiveness of our approaches.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics – Models; I.2.6 [**Computing Methodologies**]: Artificial Intelligence – Learning; J.3 [**Computer Applications**]: Life and Medical Sciences – health, medical information systems

General Terms

Algorithms, Experimentation

Keywords

Clinical decision support, health care data mining, diabetes complication, treatment, symptom

1. INTRODUCTION

Diabetes is notoriously known as a prevalent chronic disease. Till 2012, more than 371 million people have diabetes throughout the world, which results in 471 billion USD spent on healthcare for diabetes.[1] It was also reported that between 6% and 9% of North American adults had diabetes [1, 3] and were at risk for diabetes-related complications, e.g., hypertension, coronary heart disease, depression, etc. In China, diabetes had also reached epidemic proportions in the general adult population: 92.4 million adults (9.7% of the adult population) by 2010 [27]. Moreover, many complications often afflict diabetes patients seriously and incur higher healthcare expenditures. An estimated one-third of the direct medical cost of diabetes can be attributed to the management of complications [23].

Therefore, it is increasingly important to study the underlying patterns between diabetes and their complications so as to make better management and strategies aimed at the prevention, detection, and treatment of diabetes in the whole population. Previously, it is difficult to perform the study due to the lack of the sufficiently clinical data. Recently, with the fast development of hospital information systems, a large collection of electronic clinical records become available, which provides the opportunity to study medical cases, evidences and knowledge. This can benefit many real applications. For example, the inexperienced trainee medical students would like to find similar cases to learn previous evidences. Uncovering the medical knowledge and patterns hidden in the abundant medical records would also help physicians get deeper understanding on their specialty and make better decisions for the care of individual patients. Mining the medical data has attracted lots of interests from both industrial and academic communities. IBM researchers have designed a platform for intelligent care delivery analytics. In order to search similar medical cases, they proposed supervised patient similarity measures based on heterogeneous patient records [24] and introduced experts feedback into distance metric assessment [26]. Davis et al. exploited machine learning methods to predict the efficacy and risk of drugs for heart attacks [5]. Neuvirth et al. [20] studied the personalized care management of diabetes patients at risk. However, none of the previous work studied the association patterns between diabetes complications and their corresponding treatments.

Motivation In this paper, we explore the problem of mining diabetes complication and treatment patterns from a large volume of electronic medical records. Figure 1 gives an example of the medical record. We see that the medical record contains heterogeneous information, including the patient demographics, laboratory test results, radiology reports, and also some physician notes such as diagnosis, treatments, and medications. The different aspects of medical information are highly correlated and physicians are very interested in the association patterns. E.g., how to figure out patient

[1]http://www.idf.org/diabetesatlas/

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505549.

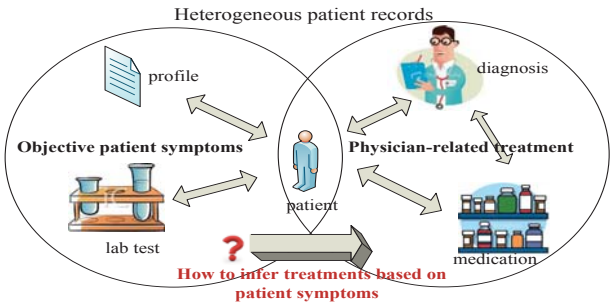


Figure 1: An example of patient medical record, which includes typical symptoms (e.g., gender, age, lab tests) and possible treatments applied to the patient. The arrows demonstrate the correlations between the heterogenous factors.

conditions from their lab test results? Which complications would often co-afflict diabetes patients with a large probability? What kind of medications should be given to the diabetes patients with some specific complications? How will the different complications afflict the different gender and age groups? Physicians always figure out these problems based on their experiences. Fortunately, the emerging electronic medical records nowadays provide us the opportunities to learn from previous cases. Efficient models can be exploited to mine these association patterns to support clinical decision-making.

Contributions To address these problems, we propose a probabilistic graphical model named Symptom-Diagnosis-Treatment model to mine patterns from the heterogeneous medical records. We use a latent variable to denote patient conditions, which the patient symptoms are assumed to be generated from. Moreover, we assume that patients' treatments are generated based on both physicians' diagnosis and patient conditions. In this way, diabetes complication and treatment patterns can be discovered simultaneously. We further study the demographics statistics of patient population w.r.t. complication patterns and observe several interesting phenomena. Finally, we propose a possible medical application in terms of treatment suggestion for clinical decision support. We test the proposed model on a collection of one-year diabetes clinical records from a famous geriatric hospital in China and our experimental results demonstrate the effectiveness of our methods.

To summarize, this work contributes on the follow aspects:

- We formally formulate the problem of diabetes complication and treatment pattern mining and propose a probabilistic generative model to simultaneously model patient symptoms, physicians' diagnosis, treatments from the heterogeneous medical records.
- We perform a qualitative analysis for the discovered complication and treatment patterns, which can help physicians learn from previous experiences and aid clinical services. We also discuss a possible medical application in terms of treatment suggestion based on the discovered patterns.
- We study the demographic statistics of patient populations w.r.t. different complication patterns on a one-year real medical records and observe several interesting phenomena for diabetes patients.

The rest of the paper is organized as follows: Section 2 discusses related work; Section 3 illustrates some data analysis on diabetes medical records; Section 4 formally formulates the problem; Section 5 explains the proposed model and lists the estimation equations of the model; Section 6 presents experimental results that validate the effectiveness of our methodology and Section 7 concludes this work.

2. RELATED WORK

Health care data mining In recent years, how to utilize the large scale electronic health care data to provide physicians with efficient decision support and objective evidence attracts the interests of researchers from various areas [8, 13]. Information scientists investigated the key issues of the electronic health record search engine and analytics platform for intelligent care delivery [10]. E.g., Hanauer et al. [12] studied the uncertainty terms in clinical documents and analyzed the query logs of an electronic health record search engine. Sun et al. [24, 26] worked on the supervised patient similarity measurement based on heterogeneous patient records and introduced experts feedback into distance metric assessment. Moreover, many researchers employed machine learning approaches to discover latent patterns from electronic medical records. E.g., Wang et al. [25] investigated the heterogeneous temporal clinical event pattern mining. A lot of work aimed to predict the efficacy or risk of a potential drug for a given individual. E.g., Rosen-Zvi et al. [21] studied the clinical and demographical factors for selecting anti-HIV therapies. Davis et al. [5] aimed to predict the risk of heart attack when consuming particular drugs. Deng et al. [6] developed an active learning method for personalized treatment.

Diabetes complication and treatment study Diabetes is a common chronic disease, which constitutes the leading cause of mortality all over the world. Recently, some researchers utilized health care data to study the risk factors for diabetes. E.g., Selby et al. [22] developed a prediction rule from clinical databases to identify high-risk patients in a large population with diabetes. Neuvirth et al. [20] studied the personalized care management of diabetes patients at risk. As diabetes causes various complications [9], considerable medical work has been conducted to investigate diabetes treatments for different complications, e.g., depression [15], coronary heart disease [19]. However, none of previous work investigated how to learn all the latent complication patterns and their corresponding treatments from large volume of medical records.

3. DATA COLLECTION AND ANALYSIS

We obtain a collection of real diabetes patient records from a hospital information system, containing around 170,000 patient records over one year period. The records consist of heterogeneous aspects, stored in different database tables, e.g., patient profiles, lab test results, pharmacy, physician profile, etc. The patient ID and the index of seeing doctors are utilized to link them.

As shown in Figure 1, we classify the data into two genres: objective patient symptoms and physician-related notes. The objective patient symptoms include the patient age, gender, the performed lab test names and the results. The physician-related data includes the physician ID, the physician specialty, given medication codes, dosages, procedure codes and diagnosis texts.

3.1 Physician-related Data

When diagnosing a particular diabetes patient at one time, physicians always write down several disease names in the record, e.g., hyperlipidemia, cerebrovascular, bronchitis, etc., which are complications of diabetes. Usually physicians give out the disease names according to ICD-10 [2] but not completely consistent. We

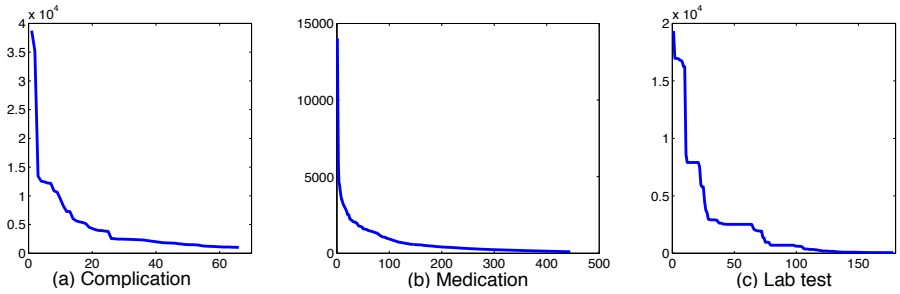


Figure 2: Diabetes complication/medication/lab test frequency

Table 1: The top 10 diabetes complications

Rank	Complication	Frequency
1	Hypertension	95,422
2	Coronary artery heart disease	54,677
3	Hyperlipidemia	38,747
4	Cerebrovascular	35,209
5	Bronchitis	13,430
6	Flu	12,537
7	Osteoporosis	12,430
8	Insomnia	12,230
9	Nephropathy	10,859
10	Chronic renal insufficiency	10,638

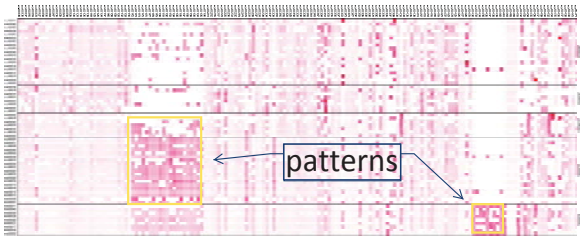


Figure 3: The co-occurrence matrix between diabetes complications and medications. Each row denotes one complication and each column denotes one medication. The darker color indicates the larger correlation weight. The gathering darker plates as marked out indicate that some association patterns indeed exist between the complications and medications.

employ some information processing methods to filter the noise of diagnosis text. Then we calculate the frequency curve of complications as shown in Figure 2(a), which demonstrates that there are about one hundred common diabetes complications in all. Table 1 shows the top 10 complications and their frequencies which are larger than 10000. We observe that the complications distribute diversely: the 5th to 10th complication respectively covers only 10% of patient records (Cf. Table 1).

The authors in work [28] classified diabetes complications into seven categories, which are cardiovascular disease, nephropathy, retinopathy, peripheral vascular disease, stroke, neuropathy, and metabolic. Different complications should be treated with different medications. The frequency curve of medications is calculated and shown in Figure 2(b). There are about 500 related medications and procedures, among which only about 50 are given in more than 2000 patient records. Figure 3 shows the co-occurrence matrix between the complications and medications, where each row denotes one complication and each column denotes one medication. We calculate their correlation weights as $\frac{N(m_i,d_j)}{\sum_{m_k} N(m_k,d_j)}$, where $N(m_i,d_j)$ represents the number of medical records which contain medication m_i and disease d_j. And we use Cluto [2] to cluster the complications and medications. The result is shown in Figure 3, where the darker color indicates the larger weight. Therefore the gathering darker plates as marked out in the figure indicate that some association patterns indeed exist between the complications and medications. In other words, particular treatments will be given according to the genre of diabetes complications.

3.2 Patient Symptoms

We analyze how three categories of diabetes complications: nephropathy, retinopathy and liver damage afflict different groups of patients w.r.t. age and gender factors. Figure 4 shows the number of patient records changing with age, where the dotted curve denotes female and the solid one denotes male. Thus we can find that generally 50 to 60 is the start bursting age span for all types of diabetes complications. The number of patients with nephropathy complications is the largest, where the male patients are much more than female ones. But for retinopathy complications the number of the female patients is larger than the male one. For the liver complications, male patients burst earlier than female ones. Therefore, different types of complications influence diabetes patient population differently w.r.t. demographics factors, i.e., age and gender.

Besides the patient demographics, lab test results are the most important information for physicians' reference to make diagnosis and treatments. From the whole record set, we count that there are about 200 types of related lab tests for diabetes patients in all. Figure 2(c) shows the frequency curve of lab tests, where the top 10 performed ones are glucose, creatinine and urea, etc., as shown in Table 2. Thus the most frequent lab tests are only performed less than 20000 times in one year, covering about 10% patient records. The reason is that 80% patients did not perform any lab tests when they saw doctors, and the rest patients performed only a small portion of all the lab tests. Thus this lab test information is very sparse, which leads to the small overlapping features between the patients even under the similar conditions. Therefore it is very difficult to measure the patient similarity only based on the lab test results.

The lab tests can have numerical or categorical types of results. We analyze the numerical results of lab tests and find that the values themselves or their transformed values approximately fit the normal distributions. For example, the left figure in Figure 5 shows the distribution of lab test "Alanine aminotransferase" result distribution, and the right one shows their log value distribution.

Although it is difficult to measure the similarity between concrete patients based on the sparse lab test information, we can get

[2] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

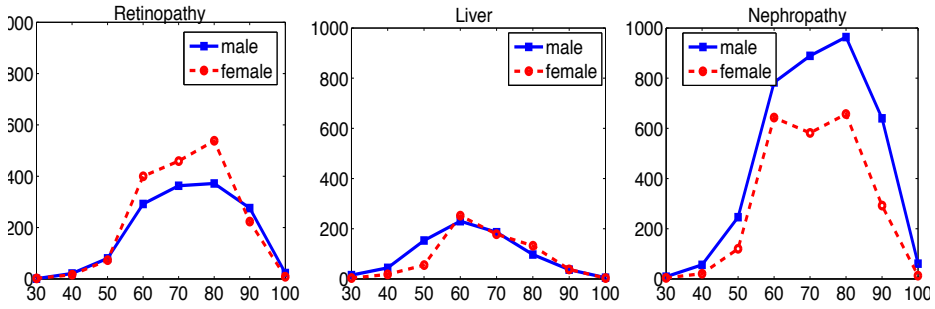


Figure 4: The size of diabetes patient population with retinopathy, liver damage and nephropathy changing with age and gender

Table 2: The top 10 lab test results statistics

Rank	Lab Test	Freq	Liver (4415)			Retinopathy (7958)			Nephropathy (23086)		
			Freq	Mean	Var	Freq	Mean	Var	Freq	Mean	Var
1	Glucose	19,342	629	1.9193	0.2798	335	1.9239	0.2642	3141	1.8753	0.2645
2	Creatinine	16,965	599	4.1213	0.2791	325	4.2138	0.3790	3112	**4.5587**	0.5869
3	Urea	16,965	599	1.6828	0.2863	325	1.7993	0.3420	3112	**2.0060**	0.4592
4	Alanine aminotransferase	16,942	623	**3.2838**	0.6733	327	2.8581	0.5529	3011	2.8117	0.5304
5	Aspartate aminotransferase	16,934	620	**3.2294**	0.4522	326	3.0076	0.3877	3010	2.9866	0.3414
6	Triglyceride	16,801	605	**0.5589**	0.6267	325	**0.3576**	0.5130	2999	**0.4496**	0.5438
7	Total cholesterol	16,792	605	1.5679	0.2173	325	1.5391	0.1879	2999	1.5197	0.2224
8	Uric acid	16,674	596	5.6614	0.3027	326	5.6672	0.3393	3072	**5.7740**	0.2890
9	Low-density lipoprotein	16,243	600	0.9809	0.3361	307	0.9876	0.2824	2779	0.9599	0.3070
10	High-density lipoprotein	16,238	598	0.2974	0.2670	307	**0.3512**	0.2268	2777	0.2785	0.2524

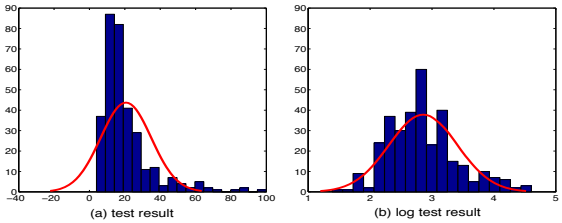


Figure 5: Alanine aminotransferase test result distribution

the statistical information from the whole data set. We calculate the log values of the frequency, mean and variation for all the numerical lab test results in the above three categories of diabetes complications respectively. Table 2 shows the results for the top 10 lab tests. The total numbers of patient records belonging to these categories of diabetes complications are also shown in the first rows, which are 4415, 7958, 23086 respectively. Therefore, the patients with retinopathy complications performed lab tests less frequently than the ones with nephropathy or liver complications. Furthermore, some lab test results do not have significant differences w.r.t. different complications, e.g., glucose, low-density lipoprotein. But some could be utilized to distinguish the complication genres, e.g., creatinine for nephropathy, triglyceride for liver complications (as the bold-faced numbers in Table 2).

Summary In this section, we analyze the collected outpatient medical records in a hospital information system over one year duration. And we conclude as the following:

- Complication and treatment patterns indeed exist in large volume of medical data as shown in Figure 3. Particular treatments should be given according to the genre of diabetes complications.
- Different types of complications will influence diabetes patient population differently w.r.t. demographics factors, i.e., age and gender, as Figure 4.
- Numerical lab test results approximately fit the normal distributions and can be used to distinguish the patient condition with different complications as shown in Table 2.

Therefore, it is feasible to explore data mining methods to study the diabetes complication patterns and to unveil the latent association mechanism between treatments and symptoms from large volume of electronic medical records. Moreover, it is essential to study the demographic statistics of patient population w.r.t. complication patterns for clinical decision support.

4. PROBLEM DEFINITION

In this section, we introduce and define several related concepts and formulate the complication and treatment pattern mining problem from heterogeneous medical records.

4.1 Definitions

As we analyzed in the above section, there are about 100 common diabetes complications. The diabetes complications co-occur among the physicians' diagnosis records and are treated similarly. It means association patterns exist between these complications. The work in [28] classified the diabetes complications into seven categories. In order to model the association patterns among complications, we define a new concept "Complication Pattern" first.

DEFINITION 1. ***[Complication Pattern]*** *A complication pattern c is defined as a spectrum of diseases d associated with weights $p(c|d)$.*

The conditional probability $p(c|d)$ represents the probability of disease d belonging to complication pattern c, based on which diseases can be classified to different complication patterns. Therefore, a complication pattern is a category of diseases which often co-occur and are treated similarly in populations. Then we define the concept of patient condition as below.

DEFINITION 2. ***[Patient Condition]*** *The patient condition θ denotes one particular patient's current condition, which is defined as a mixture of weighted complication patterns. The weight for the complication pattern c is defined as the probability with which the patient will have this complication pattern.*

The patient condition is a latent unobservable variable represented as a multinomial distribution over the complication patterns, which is the most important information physicians aim to figure out for treatments during diagnosing.

DEFINITION 3. ***[Patient Symptoms]*** *Patient symptoms are observable variables, including patient demographics, lab test results, etc., which can have categorial or numerical values.*

Intuitively patient symptoms are dependent on patient condition. As shown in Figure 5, we can assume the numerical values are drawn from some normal distributions. The parameters of distributions, i.e., the mean and variation, change with complication patterns. For categorial variables, we assume that they are generated from a multinomial probability distribution w.r.t. complication patterns. At last, we define a physician treatment.

DEFINITION 4. ***[Treatment]*** *A treatment given by a physician is defined as a mixture of medications and procedures.*

4.2 Intuitive Insights

Intuitively, the complications belonging to the same category will co-occur frequently, i.e., patients are more likely to have the complications simultaneously. For example, patients with depression are prone to have insomnia at the same time. On the other hand, they will be treated with similar therapies, i.e., patients are given some common medications. Therefore, the complication patterns can be discovered from the physicians' diagnosis and treatments.

Furthermore, patient symptoms reflect patient current conditions. Thus physicians refer to patient symptoms to figure out patient conditions. Then they give their diagnosis, based on which they treat patients. Thus patient conditions link patient symptoms and physicians' diagnosis and treatments together.

Our objective is to design a method to link different aspects of the patient records and learn the complication and treatment patterns simultaneously.

5. COMPLICATION AND TREATMENT PATTERN MINING

In this paper, we propose a probabilistic graphical model named Symptom-Diagnosis-Treatment model to mine the diabetes complication and treatment patterns from large volume of medical records. The model integrates patient symptoms and physicians' treatments, diagnosis in an unified framework so as to mine the latent correlation between complications as well as patient symptoms and physicians' treatments.

5.1 Assumptions

First we have the following two assumptions.

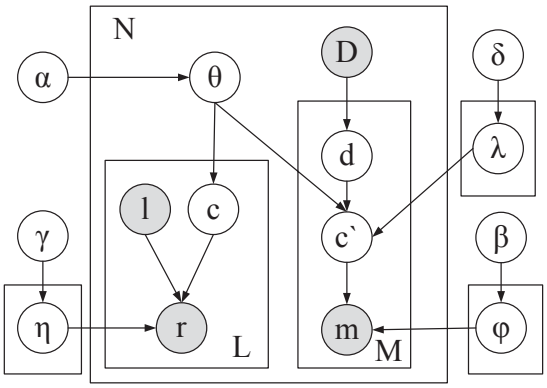


Figure 6: Symptom-Diagnosis-Treatment model

ASSUMPTION 1. *Patient symptoms, e.g., the lab test results, are generated based on patient conditions.*

As defined, patient conditions are latent and unobservable variables, which physicians aim to figure out via observing the patient symptoms. Thus it is intuitive to assume that patient symptoms are generated from patient conditions. In this paper, we mainly utilize the lab test results as patient symptoms.

ASSUMPTION 2. *Physicians give out treatments based on both their diagnosis and patient conditions.*

After observing patient symptoms, physicians will have their own opinions (the diagnosis) on patient conditions and give their corresponding treatments, e.g., medications and procedures. In real situation, physicians diagnose patient conditions based on their experiences. Thus even for the same patient with the same condition, different physicians may have different opinions so as to give out different treatments. Therefore we should consider the subjective bias of physicians in medical records. And different from patient symptoms in Assumption 1, the treatments are dependent on both the diagnosis and patient conditions.

5.2 Probabilistic Generative Model

Based on the assumptions, we propose the Symptom-Diagnosis-Treatment model as shown in Figure 6. The variables are described in Table 3. In the model, θ represents the patient condition which is generated from a Dirichlet prior α. The left part represents the generation process of patient symptoms. For each patient symptom, first a complication pattern c is selected based on the patient condition θ. Then the patient symptom is generated either based on c-specific normal distribution if its values are numerical or from c-specific multinomial distribution if its values are categorical. The right part represents the generation process of medications, which is performed in a similar way. The parameter λ indicates the weights of diseases for complication patterns as Definition 1. The observed variable D indicates the physician's diagnosis for the particular medical record. The treatment is generated based on both the patient condition and the diagnosis. Thus a diagnosed complication d is first generated from D, and then a complication pattern c' is generated based on the diagnosed complication d and patient condition θ. The whole process is shown in Alg. 1.

5.3 Model Estimation via Gibbs Sampling

In this paper, we use Gibbs sampling method to estimate the model as we assume conjugate prior for the parameters θ,ϕ, λ and

Table 3: Variable descriptions

Notation	Description
c, c'	complication pattern assigned to patient symptoms and physician treatments
l	the patient symptom type, e.g., the lab test type
r	the patient symptom result, .e.g., the lab test result
m	the given medication or precedure
D	the complication set of physician diagnosis
d	the selected complication
N	the number of patient records
M	the number of medications and precedures
L	the number of patient symptoms
K	the number of complication patterns
θ	the mixture of complication patterns which indicate the patient condition
α	the Dirichlet prior for patient condition θ
η	the distribution parameters for patient symptoms
γ	the prior for η
λ	the mixture of complication patterns for each disease
δ	the Dirichlet prior for λ
ϕ	treatment distributions w.r.t. complication patterns
β	the Dirichlet prior for ϕ

η [7]. Gibbs sampling is an algorithm to approximate the joint distribution of multiple variables by drawing a sequence of samples, which iteratively updates each latent variable under the condition of fixing remaining variables [11]. We utilize the function $N()$ to store the number of samples during Gibbs sampling. For example, $N_{x,c'}(x, c')$ represents the number of samples in patient record x which are assigned to pattern c'. $N_{d,c'}(d, c')$ represents the sample number of disease d which are assigned to pattern c'. $N_{m,c'}(m_i, c')$ represents the sample number of medication m_i which are assigned to pattern c'. Besides, we use $*$ to represent the sum operation. E.g., $N_{m,c'}(*, c')$ represents the sum number of all medication samples which are assigned to pattern c', i.e., $N_{m,c'}(*, c') = \sum_i N_{m,c'}(m_i, c')$. Thus the update estimation equations are shown as below.

$$p(c'_i|\vec{c}'_{-i}, m_i, .) \propto \frac{N_{x,c'}(x_i,c'_i)+N_{x,c}(x_i,c'_i)+N_{d,c'}(d_i,c'_i)+\alpha+\phi}{N_{x,c'}(x_i,*)+N_{x,c}(x_i,*)+N_{d,c'}(d_i,*)+K\cdot\alpha+K\cdot\phi} \cdot \frac{N_{m,c'}(m_i,c'_i)+\beta}{N_{m,c'}(*,c'_i)+M\cdot\beta} \quad (1)$$

$$p(c_j|\vec{c}_{-j}, r_j, .) \propto \frac{N_{x,c'}(x_j,c_j)+N_{x,c}(x_j,c_i)+\alpha}{N_{x,c'}(x_j,*)+N_{x,c}(x_j,*)+K\cdot\alpha} \cdot \frac{N_{c,r,l}(c_j,r_j,l_j)+\gamma}{N_{c,r,l}(*,r_j,l_j)+N_l\cdot\gamma} \quad (2)$$

$$p(c_j|\vec{c}_{-j}, r_j, .) \propto \frac{N_{x,c'}(x_j,c_j)+N_{x,c}(x_j,c_i)+\alpha}{N_{x,c'}(x_j,*)+N_{x,c}(x_j,*)+K\cdot\alpha} \cdot Norm(r_j|\mu'_0, \sigma'_0) \quad (3)$$

where

$$\mu'_0 = \sigma'_0 \cdot (\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{N_{l,c}(l_j,c_j)} r_i}{\sigma^2})$$
$$\sigma'_0 = (\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2})^{-1} \quad (4)$$

For the symptoms with categorical values, the update equation is Equ. (2), where N_l denotes the number of possible values for patient symptom l. For the symptoms with numerical values, we assume they are generated from normal distributions with known variations σ, whose mean value μ is generated from a normal distribution prior, i.e., $Norm(\mu|\mu_0, \sigma_0)$. As normal distribution's conjugate prior is normal distribution, we can get Equ. (3) and (4) [18].

5.4 Discussion

It is very efficient and convenient to employ probabilistic graphical models to discover the latent patterns in heterogeneous networks [4, 16, 17]. Researchers can design other types of graphical models based on their definitions of latent variables and assumptions of the dependence between the variables on this problem. For example, diagnostics can be deemed as a sequential process derived from the patient symptoms so they are assumed to be generated from patient conditions as well. They can also be classified as true or false ones which are supposed to result in different medications. The factor of physicians can be included in the model so that we can assign different trust weights to the diagnostics, and etc.

In our implementation process, we have tried several forms of probabilistic graphical models based on the above ideas. In this paper we report the model in Figure 6 which has the best experimental results among all we have tried. But we still think there exists improvement room for this model. And we can further investigate them according to different applications later.

foreach *patient record x* **do**
  foreach *medicine i in given medication set* **do**
    Draw a disease d_i from the diagnosis set D ;
    Draw a complication pattern $c'_{x,i}$ based on both the diagnosis d_i and the patient condition θ_x, i.e., $c'_{x,i} \sim multi(\lambda_{d_i} + \theta_x)$;
    Draw a medicine m_i from $c'_{x,i}$-specific medicine distribution, i.e., $m_i \sim multi(\phi_{c'_{x,i}})$;
  end
  foreach *lab test j in done lab test set* **do**
    Draw a complication pattern $c_{x,j}$ from the patient condition θ_x, i.e., $c_{x,j} \sim multi(\theta_x)$;
    if *l is numerical lab test* **then**
      Draw a lab test result r_j from $c_{x,j}$-specific normal distribution for lab test l's results, i.e., $r_j \sim norm(\eta_{l,c_{x,j}})$;
    end
    if *l is categorical lab test* **then**
      Draw a lab test result r_j from $c_{x,j}$-specific multinomial distribution for lab test l's results, i.e., $r_j \sim multi(\eta_{l,c_{x,j}})$;
    end
  end
end

Algorithm 1: Probabilistic generative process of Symptom-Diagnosis-Treatment model

6. EXPERIMENTS

In this section, we present experimental results on real diabetes medical records to evaluate the efficiency and effectiveness of the proposed approach.

6.1 Experimental Setup

We collect about 177,000 real electronic medical records of diabetes over one year duration from a famous geriatric hospital in Beijing, China, on which our model is employed to discover diabetes complication and treatment patterns.

Data preprocess First, we preprocess the medical records and filter out some noisy data. As shown in Figure 2, there is a large proportion of medicines, lab tests and complications which only appear very few times. Thus we remove the unfrequent items to reduce the influence of noise. On the other hand, some lab tests are not distinguished, which almost have the same result in all the records. Therefore we utilize the information entropy to measure the effects of the lab tests and eliminate the undistinguished ones. In this way, we get a data set as shown in Table 4.

Table 5: Complication and treatment patterns. Each column indicates one complication and treatment pattern, which is represented by several most related diseases and medications and their probabilities.

Nephropathy		Inflammation		Osteoporosis	
disease	prob	disease	prob	disease	prob
Chronic renal insufficiency	0.957	Bronchitis	0.952	Severe osteoporosis	0.946
Hyperuricemia	0.930	Tracheitis	0.917	Bone pain	0.941
Hypothyroidism	0.863	Flu	0.897	Prostatic hypertrophy	0.895
Gout	0.435	Cold	0.895	Osteoporosis	0.812
Urinary tract infection	0.775	Upper respiratory tract infection	0.874	Gastrointestinal disorders	0.801
Diabetic nephropathy	0.735	Chronic bronchitis	0.851	Severe osteoarthritis	0.790
medicine	prob	medicine	prob	medicine	prob
Cordyceps	0.132	Acarbose tablets	0.069	Calcium carbonate tablets	0.109
Sodium bicarbonate tablets	0.037	Cold heat particles	0.031	Acarbose tablets	0.079
Benzbromarone	0.032	Aspirin	0.029	Calcitriol capsules	0.065
Depression		**Cardiovascular**		**Peripheral neuropathy**	
disease	prob	disease	prob	disease	prob
Depression	0.750	Hyperlipidemia	0.904	Peripheral neuropathy	0.756
Depressive state	0.700	Atherosclerosis	0.882	Neuritis	0.689
Insomnia	0.690	High cholesterol	0.876	Neurosis	0.687
Neurasthenia	0.653	Arteriosclerosis	0.847	Peripheral nerve damage	0.571
Neurosis	0.629	Arrhythmia	0.527	Vitamin deficiency	0.549
		Gout	0.811	Diabetic peripheral neuropathy	0.504
medicine	prob	medicine	prob	medicine	prob
Estazolam tablets	0.090	Aspirin	0.077	Vitamin B1	0.192
Acarbose tablets	0.069	Acarbose tablets	0.062	Glucosamine hydrochloride	0.052
Aspirin	0.031	Atorvastatin calcium	0.052	Methylcobalamin	0.042
Digestion disease		**Retinopathy**		**Liver disease**	
disease	prob	disease	prob	disease	prob
Constipation	0.921	Diabetic retinopathy	0.888	Fatty liver	0.586
Indigestion	0.894	Cataract	0.836	Liver damage	0.512
Gastric ulcer	0.851	Type 2 diabetic neuropathy	0.835	Elevated transaminase	0.418
Lumbago	0.787	Diabetic neuropathy	0.827	Hypertension	0.406
Peptic ulcer	0.706	Conjunctivitis	0.736	High cholesterol	0.112
medicine	prob	medicine	prob	medicine	prob
Painkiller plaster	0.103	Methylcobalamin tablets	0.063	Metformin hydrochloride tablets	0.143
Acarbose tablets	0.084	Pancreatic kallikrein	0.055	Pioglitazone hydrochloride tablets	0.142
Glucosamine Hydrochloride	0.052	Vitamin B1	0.054	Acarbose tablets	0.140

Table 4: Data set

	Maximum frequency	size in selected set
complication	38747	70
medication	14025	550
lab test	19342	70

Evaluation Aspects Then, we apply our model to the whole data and evaluate our method on the following aspects:

- **Complication pattern illustration** We illustrate the complication patterns discovered from our model, from which we find some interesting correlations among the 70 diabetes complications. These complication patterns demonstrate the major categories of diabetes complications in one-year medical records in China. We compare them with the diabetes complication categories of US proposed in previous work [28].
- **Treatment pattern demonstration** Besides complication patterns, our model also discovers treatment patterns. The results demonstrate that according to different types of diabetes complications the treatments should be different. But there also exist some common medications for several complication patterns. Furthermore, our model discovers the correlations between patient symptoms and conditions. We show some symptoms can be utilized to distinguish patient conditions, i.e., which types of complications the patient may have.
- **Complication pattern statistics** We analyze the characteristics of complication patterns for patient demographics. E.g., how big is the patient population with each complication pattern? What are their age trends? Will complications afflict different gender and age group differently? Furthermore, we study the impact of complication combinations on patient conditions. E.g., are older people prone to have more complications? Will complication combinations influence the severity of patient conditions? We do all these statistical analysis and get some interesting results.
- **Clinical decision support case study** The complication and treatment patterns discovered from our model can be utilized to support clinical decision-making and aid clinical diagnosis. We show a medical application of suggesting possible medications according to the lab test results of particular patients. In order to test the model quantitatively, we use MAP, Precsion@10, Recall@10 to measure the effects of the application and compare our approach with the nearest neighbor model.

6.2 Complication and Treatment Patterns

Table 5 visualizes the complication and treatment patterns obtained from the Symptom-Diagnosis-Treatment model when we set the number of patterns to be 9. Each column indicates one complication and treatment pattern, which is represented by several most related diseases and medications and their probabilities. Therefore we can tell the meanings of the complication patterns and summarize them in the captions. So the discovered diabetes complication

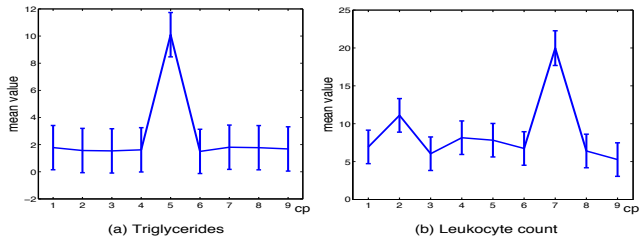


Figure 7: Lab test results distribution on the nine complication patterns ((a) triglycerides (b) leukocyte count)

patterns are respectively associated with "nephropathy", "inflammation", "osteoporosis", "depression", "cardiovascular", etc.

Complication pattern demonstration The complication patterns show the latent correlations among the diabetes complications, from which we can find some interesting phenomena. For example, the third complication pattern shows that there exist latent correlations among the diseases "osteoporosis", "prostatic hypertrophy" and "gastric disease" although they are related to different human organs. The fourth complication pattern shows that there is a large probability for patients with "depression" to have "insomnia" at the same time.

Young et al. [28] studied seven categories of diabetes complications in US, which are cardiovascular disease, nephropathy, retinopathy, peripheral vascular disease, stroke, neuropathy, and metabolic. Our model discovers some consistent categories on the medical records in China, e.g., nephropathy, retinopathy, cardiovascular, neuropathy. But we also discover some other different categories of diabetes complications, e.g., depression, which have also been studied a lot in many works [15].

Treatment pattern demonstration Our model can also discover symptom and treatment patterns w.r.t. different types of complications.

As Definition 4, the treatment patterns are represented as the conditional probabilities $p(m|c)$. Table 5 also shows the top 3 relative medications to each complication pattern [3]. Therefore there obviously exist different treatments w.r.t. different types of diabetes complications. Meanwhile, there also exist some common medications for several complication patterns, e.g., acarbose tablets. It means this kind of medication is a widely used drug for diabetes.

Patient symptoms distribution Different types of diabetes will result in different symptoms. Thus patients are required to do some lab tests to provide physicians with the references for clinical diagnosis. Our models can discover the distribution of the lab test results on the complication patterns. For example, Figure 7 shows the mean and variation values of the lab test "triglycerides" and "leukocyte count" on the nine complication patterns, which demonstrates that triglycerides has a large value associated with complication pattern 5. That is because "hyperlipidemia"-related complications lead to much higher "triglycerides" test results. Conversely, if a patient has a high "triglycerides" test result, physicians will diagnose he/she has hyperlipidemia. The value of "Leukocyte count" is large on complication patterns 2 and 7, which is caused by "inflammation"-related complications. Thus large "leukocyte count" indicates the existence of "inflammation" complications.

[3]All the disease and medication names are translated by http://translate.google.com/

6.3 Patient Statistics on Complication Patterns

We would like to see how these discovered complication patterns distribute in diabetes patient population. We classify the patients into nine clusters by checking their diagnosis. If the diagnosis contains any disease belonging to one complication pattern, we assume the patient has this complication pattern. Thus, some patients will belong to more than one cluster.

Patient statistics w.r.t. complication patterns

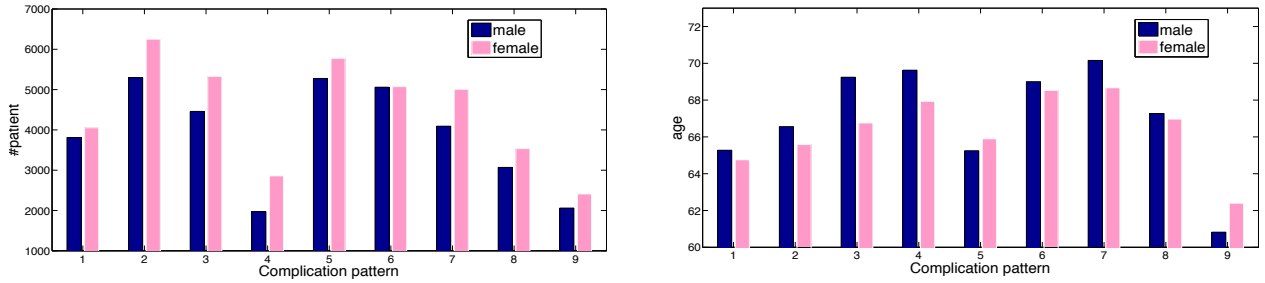
First, we want to study the characteristics of individual complication patterns w.r.t. patient demographics. E.g., what is the size and average age of patient population with each complication pattern, particularly for men and women? We count the number of male and female patients in these clusters, which is shown in Figure 8(a). We also calculate their average ages as shown in Figure 8(b). Therefore, we can find the following phenomena.

- The groups of patients with complication patterns "depression" and "liver disease" are smaller than other groups. Contrastively, patients are more likely to have the diabetes complications "inflammation" and "cardiovascular".
- The patients with liver complications (i.e., complication pattern 9) are obviously the youngest. In other words, the diabetes patients are likely to get liver complications at an earlier age than other types.
- With all kinds of complication patterns, female diabetes patient population is generally larger than male population, which is notable in the groups of the complication pattern "depression" and "osteoporosis". Thus women are easier to get these types of complications than men if they have diabetes.
- In most cases, female patients have the diabetes complications earlier than male ones. But liver complication pattern is one counter-example, i.e., male patients get liver complications at a smaller age than female ones. The reason may be that men drink alcohol more than women.

Patient statistics w.r.t. complication pattern combinations

In many real cases, diabetes patients are probably to have more than one type of complications [28], i.e., they will belong to several diabetes groups simultaneously. We further study the impact of complication pattern combinations on patient conditions and find the following results:

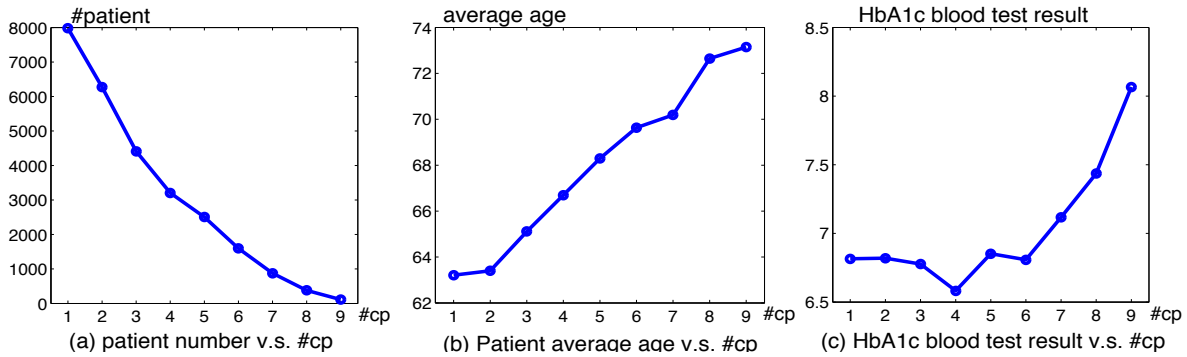
- Figure 9(a) shows the curve of the number of patients who have 1 to 9 complication patterns, which indicates that patient population declines with the size of complication pattern combinations increasing. In other words, patient population with more complication patterns is smaller than the one with fewer complication patterns.
- Figure 9(b) shows the average age curve changing with the size of complication pattern combinations, which demonstrate the average age has an obvious growth trend when the size of complication pattern combinations increases. Thus we conclude that the older patients are likely to have more types of complication patterns simultaneously than younger ones.
- The HbA1c blood test (also called glycohemoglobin) is a reliable indicator of patient condition. The larger the HbA1c measure, the higher the risk of developing complications such as eye, heart, kidney disease, nerve damage or stroke [20]. Thus we calculate the average values of HbA1c test results of the above patient groups as shown in Figure 9(c), the increasing



(a) Patient population w.r.t. complication patterns

(b) Patient average ages w.r.t. complication patterns

Figure 8: Patient statistics w.r.t. the nine complication patterns



(a) patient number v.s. #cp

(b) Patient average age v.s. #cp

(c) HbA1c blood test result v.s. #cp

Figure 9: Statistics on patients who have several complication patterns simultaneously (#cp represents the number of complication patterns)

trend in which tells us that the patients with more types of complication patterns are probably at higher risk in general.

6.4 Application Example

Research on clinical decision support [14] has developed for many years, which both physicians and patients pay much attention to. E.g., when patient symptoms are observed, physicians want to figure out these issues: what kinds of diabetes does this patient have? which medications should be given to him/her? The complication and treatment patterns discovered by our model can be used to help physicians learn from previous experiences so as to aid their work. Here, we propose one possible medical application based on our model.

As illustrated above, the treatment probability $p(m|c)$ as well as each symptom's distribution w.r.t. complication $p(r|c,l)$ are discovered from the model. As $p(c|r,l) \propto p(r|c,l) \cdot p(c)$, the most likely medications can be obtained by the following equation.

$$p(m) \propto \sum_{c} p(m|c) \cdot p(c|r,l) \tag{5}$$

In order to test our model quantitatively, we do some experiments based on this application schema. Suppose the diagnosis and treatments given by physicians in the medical records are the ground-truth (In real situations, the given diagnosis and treatments are biased. Even for the same patient, different physicians may have different opinions on patient conditions so as to give different diagnosis and treatments.). Then we test the consistency of the possible medications and diseases suggested by our model with the ground-truth. We randomly select 10000 medical records which contain lab test results and diagnosis as testing data and use the rest records to train the model. We utilize three measurements which are mean average precision (MAP), mean precision at top 10 (MP@10) and mean recall at top 10 (MR@10). We compare the results with k-nearest neighbor model (NNM), i.e., using the weighted combination of medications of the k nearest neighbors as the suggested treatment (we use $k = 10$ here). The results in Table 6 demonstrate that our model can outperform the k-nearest neighbor model in terms of the medicine suggestion application.

Table 6: Medical Application Result

model	MAP	MP@10	MR@10
our model	0.541	0.580	0.476
NNM	0.121	0.143	0.098

7. CONCLUSION AND FUTURE WORK

In this paper, we explore the problems of using large volume of electronic medical records to aid the clinical treatments of diabetes mellitus. We propose a probabilistic graphical model to link the patient symptoms and physicians' diagnosis, treatment together and to mine diabetes complication and treatment patterns simultaneously. We further study the demographic statistics of patient population w.r.t. complication patterns in real data and discover some interesting insights. The experimental results on a collection of one-year diabetes clinical records from a famous geriatric hospital in Beijing, China demonstrate the meaningful patterns discovered by our models.

The general problem of exploring electronic medical records for clinical decision support represents a promising and interesting research direction in health care data mining. There are many potential applications of this work. For example, chemists want to find

symptom-treatment patterns from abundant records of traditional Chinese medicine so as to develop new drugs. Therefore one potential future work is to apply the model to other genres of medical data, e.g., traditional Chinese medicine records.

8. ACKNOWLEDGMENTS

This work is supported in part by the following grants: U.S. NSF awards CCF-0833131, CNS-0830927, IIS-0905205, CCF-0938000, CCF-1029166, and OCI-1144061; DOE awards DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, DESC0005340, DESC0007456; AFOSR award FA9550-12-1-0458 and National Natural Science Foundation of China under grants 61103065. Jie Tang is supported by Natural Science Foundation of China (No. 61222212, No. 61073073), and a research fund supported by Huawei Inc.

9. REFERENCES

[1] Report from the national diabetes surveillance system: Diabetes in canada, 2009. Technical report, Public Health Agency of Canada, 2009.

[2] International classification of diseases (icd). Technical report, World Health Organization, 2010.

[3] National diabetes fact sheet: National estimates and general information on diabetes and prediabetes in the united states, 2011. Technical report, Atlanta, GA, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2011.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[5] J. Davis, E. Lantz, D. Page, J. Struyf, P. Peissig, H. Vidaillet, and M. Caldwell. Machine learning for personalized medicine: Will this drug give me a heart attack? In *Machine Learning in Health Care Applications Workshop. In conjunction with ICML 2008*, 2008.

[6] K. Deng, J. Pineau, and S. A. Murphy. Active learning for developing personalized treatment. In *UAI'11*, pages 161–168, 2011.

[7] A. Doucet, N. de Freitas, K. Murphy, and S. Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. In *UAI'00*, pages 176–183, 2000.

[8] S. Ebadollahi, A. R. Coden, M. A. Tanenblatt, S.-F. Chang, T. Syeda-Mahmood, and A. Amir. Concept-based electronic health records: opportunities and challenges. In *MULTIMEDIA '06*, pages 997–1006, 2006.

[9] C. for Disease Control Diabetes in Managed Care Work Group. Diabetes mellitus in managed care: complications and resource utilization. *Am J Manag Care*, 7:501–508, 2001.

[10] D. Gotz, H. Starvropoulos, J. Sun, and F. Wang. Icda: A platform for intelligent care delivery analytics. In *American Medical Informatics Association Annual Symposium (AMIA)*, 2012.

[11] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

[12] D. A. Hanauer, Y. Liu, Q. Mei, F. J. Manion, U. J. Balis, and K. Zheng. Hedging their mets: The use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. In *American Medical Informatics Association Annual Symposium (AMIA)*, 2012.

[13] J. Hu, F. Wang, J. Sun, R. Sorrentino, and S. Ebadollahi. A healthcare utilization analysis framework for hot spotting and contextual anomaly detection. In *American Medical Informatics Association Annual Symposium (AMIA)*, 2012.

[14] D. L. Hunt, R. B. Haynes, S. E. Hanna, and K. Smith. Effects of computer-based clinical decision support systems on physician performance and patient outcomes. *JAMA: The Journal of the American Medical Association*, 280(15):1339–1346, October 1998.

[15] W. Katon, M. Von Korff, E. Lin, and et al. Improving primary care treatment of depression among patients with diabetes mellitus: the design of the pathways study. *Gen Hosp Psychiatry*, 25:158–168, 2003.

[16] L. Liu, J. Tang, J. Han, and S. Yang. Learning influence from heterogeneous social networks. *Data Min. Knowl. Discov.*, 25(3):511–544, 2012.

[17] L. Liu, F. Zhu, L. Zhang, and S. Yang. A probabilistic graphical model for topic and preference discovery on social media. *Neurocomput.*, 95:78–88, Oct. 2012.

[18] K. P. Murphy. Conjugate bayesian analysis of the gaussian distribution. Technical report, 2007.

[19] S. Natarajan, Y. Liao, G. Cao, S. Lipsitz, and D. McGee. Sex differences in risk for coronary heart disease mortality associated with diabetes and established coronary heart disease. *Arch Intern Med*, 163:1735–1740, 2003.

[20] H. Neuvirth, M. Ozery-Flato, J. Hu, J. Laserson, M. S. Kohn, S. Ebadollahi, and M. Rosen-Zvi. Toward personalized care management of patients at risk: The diabetes case study. In *KDD '11*, pages 395–403, 2011.

[21] M. Rosen-Zvi, A. Altmann, M. Prosperi, E. Aharoni, H. Neuvirth, A. Sönnerborg, E. Schülter, D. Struck, Y. Peres, F. Incardona, R. Kaiser, M. Zazzi, and T. Lengauer. Selecting anti-hiv therapies based on a variety of genomic and clinical factors. *Bioinformatics*, 24(13):i399–i406, July 2008.

[22] J. Selby, A. Karter, L. Ackerson, A. Ferrara, and J. Liu. Developing a prediction rule from automated clinical databases to identify high-risk patients in a large population with diabetes. *Diabetes Care*, 24:1547–1555, 2001.

[23] S. Simpson and et al. The cost of major comorbidity in people with diabetes mellitus. *CMAJ*, 168:1661–1667, 2003.

[24] J. Sun, F. Wang, J. Hu, and S. Edabollahi. Supervised patient similarity measure of heterogeneous patient records. *SIGKDD Explor. Newsl.*, 14(1):16–24, Dec. 2012.

[25] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi. Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In *KDD '12*, pages 453–461, 2012.

[26] F. Wang, J. Sun, and S. Ebadollahi. Integrating distance metrics learned from multiple experts and its application in inter-patient similarity assessment. In *SDM'11*, pages 59–70, 2011.

[27] W. Yang and et al. Prevalence of diabetes among men and women in china. *The New England Journal of Medicine*, 362:1090–1011, March 2010.

[28] B. Young, E. Lin, M. Von Korff, G. Simon, P. Ciechanowski, E. Ludman, S. Everson-Stewart, M. Kinder, L. andO liver, E. Boyko, and W. Katon. Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization. *American Journal of Managed Care*, 14(1):15–23, 2008.