Collaborative Prediction Model of Disease Risk by Mining Electronic Health Records

Shuai Zhang, Lei Liu, Hui Li, and Lizhen Cui(✉)

School of Computer Science and Technology,
Shandong University, Jinan, China
zhangshuai0l@mail.sdu.edu.cn,
{l.liu,lih,clz}@sdu.edu.cn

Abstract. Patient Electronic Health Records (EHR) is one of the major carriers for conducting preventative medicine research. However, the heterogeneous and longitudinal properties make EHRs analysis an inherently challenge. To address this issue, this paper proposes CAPM, a Collaborative Assessment Prediction Model based on patient temporal graph representation, which relies only on a patient EHRs using ICD-10 codes to predict future disease risks. Firstly, we develop a temporal graph for each patient EHRs. Secondly, CAPM uses hybrid collaborative filtering approach to predict each patient's greatest disease risks based on their own medical history and that of similar patients. Moreover, we also calculate the onset risk with the corresponding diseases in order to take action at the earliest signs. Finally, we present experimental results on a real world EHR dataset, demonstrating that CAPM performs well at capturing future disease and its onset risks.

Keywords: Electronic Health Records · Temporal graph · Collaborative prediction · Disease risk profile

1 Introduction

Healthcare is increasingly becoming an important research field that is closely related to everyone's daily life. A huge amount of money is wasted every year due to the high degree of complexity in medical area. This crisis has motivated the drive towards preventative medicine, where the main concern is identifying the onset risk of diseases and taking preventive measures at the earliest signs [1]. Patient EHRs are systematic collections of patients' longitudinal clinical information generated from different healthcare industry institutions. Effective utilization of EHR data is the key to many medical informatics research problems [2]. Working directly with raw EHRs is very challenging due to its sparsity, noise and the existence of heterogeneity. To address this challenge, we should first do consistent representation for each patient before going into the stage of detailed disease risk prediction applications, which is a basic step to transform the raw EHRs into clinically relevant information.

Based on the EHR data, care providers typically want to assess the risk scores of a patient developing different diseases. Once the risk of a patient is predicted, proper intervention and care plan can be designed accordingly. A lot of diseases have

S. Wang and A. Zhou (Eds.): CollaborateCom 2016, LNICST 201, pp. 71–82, 2017.
DOI: 10.1007/978-3-319-59288-6_7

preventable risk factors or at least indicators of disease onset risk. Adequately describing the characteristics of these diseases may assist in preventative medicine, and help reduce the burden of disease [3]. However, it is impossible for an individual medical doctor to give a sufficient real-time analysis in the process of patient interaction, due to the complexity of risk factors' possible combination. Thus, we need a computational analysis model to take effective measures in preventive medicine. For instance, we can integrate and utilize the medical data of patients, discover deep knowledge about patient similarity relationship, and provide personalized disease risk profiles for each individual patient. The data above derived from not only the EHRs of patient, but also from similarities of the patient to thousands of other patients [4].

To deal with the aforementioned problems, this paper proposes an integrative temporal graph representation based collaborative assessment prediction model called CAPM which mainly consists of two parts: *Patient Temporal Graph* and *Disease Risk Profile*. The first part transforms temporal clinical events that extracted from raw EHRs of each patient into medical temporal graph. In open literatures, studies [2, 7, 8] proposed how to represent patient EHRs, which represents the patient historical records in sequence [2], matrix [7] and graph [8] respectively. The works above did not consider the temporal relationship between different clinical events. Thus, an approach is developed to construct the temporal graph for each patient EHRs.

The second part *Disease Risk Profile* is to predict the most probable diseases that a patient will develop in the future. Our work is inspired by learning from the work on collaborative filtering methodology [9–11] used in other settings and motivated by patient-centric model that creates a personalized healthcare [3, 4, 12]. The difference with these studies is that we utilized a *hybrid collaborative filtering approach* based on temporal graph representation to calculate the disease risk of individual patient. We calculate the similarity between a patient's record and other patients' records, and then derived the risk of a certain disease. More importantly, our CAPM also calculates the onset of the corresponding certain disease. The output is a ranked list of diseases and corresponding disease's onset risk for a patient. Thus, the patient's *disease risk profile* obtained by our method not only includes the list of diseases, but also contains the onset risk of each disease.

The main contributions of this paper are summarized as follows:

(1) The CAPM provides a unified representation (i.e., temporal graph) to express each patient's raw EHR data, and can conveniently extract ICD-10 codes in chronological order from graph to predict future disease risk.
(2) A hybrid collaborative prediction approach is developed, combining three kinds of similarity calculation methods and proposing an approach to calculate the onset risk of disease.
(3) Extensive experiments on a real-world EHR dataset are implemented to prove the predictive effectiveness of our model.

The remainder of this paper is organized as follows. Section 2 reviews related work. In Sect. 3 we describe an outline of our proposed model. The details of temporal graph based collaborative prediction model CAPM are presented in Sect. 4. Section 5 studies the performance of the proposed model through real world experiments. Section 6 concludes this paper.

2 Related Work

Patient EHR data is collected over time on patients' clinical information and is becoming section of big data revolution [3]. Furthermore, EHRs contain heterogeneous data such as diagnoses, medications lab results, and etc. Diverse modeling techniques are needed to meet the heterogeneity of EHR data, which offers many options for their combination [13]. There are a number of related works on how to represent patient EHR data [2, 7, 8]. However, these works did not consider the temporal relationships among different clinical events, which are the crucial information on the impending disease conditions.

It is a hot research topic to predict disease risks and rank diseases by their risks for individuals in data mining techniques. Some researches have been done along this line of thought. Davis *et al*. [12] proposed CARE, which is the first well-known system using collaborative filter technique to predict the disease risk of one patient. Hussein *et al*. [14] proposed Integrated Collaborative Filtering framework to develop recommender system to suggest medical advice to patients. However, all these works did not consider the onset time risk of patient's each ranked list disease.

On the one hand, this paper develops a temporal graph based representation for patient raw EHR data. Not only the temporal relationships are considered, but also the close degree of clinical events connection is computed. Thus, we can conveniently extract the required information based on the temporal graph, which make the results more interpretative. While, the existing researches are based on the raw data to conduct similarity study. On the other hand, we develop a hybrid collaborative filtering approach, which combing three kinds of similarity calculation methods. It is different from the traditional calculation using one way to compute similarity. In addition, this paper proposes an approach to calculate the onset risk of each predicted disease, so that take action at the earliest signs.

3 Overview of Collaborative Prediction Model

The preliminary of CAPM is illustrated in Fig. 1. Our model mainly contains two parts: the construction of the temporal graph from patient's raw EHR data, and the diagnoses extracted from temporal graph in time order, which are used to construct the disease risk profile of patient. The following is a brief summary of two parts.

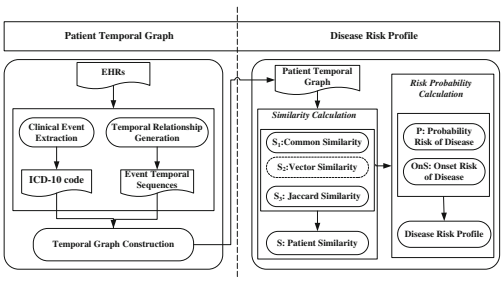

Fig. 1. The overview of CAPM

3.1 Patient Temporal Graph

This part provides a unified view for each patient by summarizing the longitudinal EHR data and from this view we can capture holistic temporal information for collaborative prediction analysis task. First, this component extracts clinical events, and generates the temporal sequences among these events based on the timestamp. Then, the obtained event sequences are transformed into the temporal graphs. This representation can use a more compact way to capture temporal structures hidden in the event sequences. Moreover, the repeated pairwise events with the same ordering in patient sequence will only appear once in patient temporal graph, which means this representation is resistant to sparse and irregular observations. Details of the temporal graph construction are described in Sect. 4.1.

3.2 Disease Risk Profile

The *Disease Risk Profile* part provides a hybrid collaborative filtering algorithm to obtain the individual patient's risk profile based on the temporal graph. The detailed EHR data documents the clinical events in time, which typically includes diagnosis, medication, and lab test. The diagnosis events are among the most structured and informative events, for which they are regarded as the prime candidates for constructing features for risk prediction. In the temporal graph, the diagnosis events are often in the form of International Classification of Diseases 10 (ICD-10) codes. Each disease is given a unique code, and can be up to 6 characters long. For example, code *I10* represents *Essential Hypertension*, *I10.X02* and *I10.X08* indicates the *Benign Hypertension* and *Hypertension* respectively. We can obtain the ICD-10 codes in time sequence based on the each patient temporal graph. Then, combine three methods to calculate the patient's similarity. Moreover, use our proposed approach to calculate the onset risk of corresponding disease. The output of this part is the disease risk profile for each patient, consisting of two aspects which are ranked list of diseases and corresponding disease's onset risk.

4 Collaborative Assessment Prediction Model

This section presents the details of the collaborative assessment prediction model to predict a ranked list of potential diseases and corresponding onset risk for a patient. In the first step, patients' medical histories are represented in the form of temporal graphs, which are constructed from the raw EHR data. After data cleaning and expressing, the patients' diagnoses are fed into the hybrid collaborative filtering approach, training it to predict comorbidities and onset risk of the disease. When the model is applied to a new patient's record, the collaborative filtering computes and selects the neighborhood of patients who are most similar to the specific patient. Finally, the likelihood of each possible disease is calculated, and a ranked list of possible diseases and its onset risk based on the likelihood is built for this patient.

4.1 Temporal Graph Representation

Inspired by Liu *et al*. [5], we construct the following temporal graph for each patient's sequence s_n:

Definition 1 ***(Temporal Graph).*** Let temporal graph G_n of sequence s_n be a directed and weighted graph $G = (V, E)$ with vertex set V and directed edge set E. The weight of the edge from node i to node j is defined as the averaged temporal closeness between any *ij-th* of each input event sequence s_n:

$$W_{ij}^n = \sum_{1 \le p \le q \le L_m} [e_p^n = i \wedge e_q^n = j] \frac{1}{L_m} \delta_\mu (t_q^n - t_p^n). \tag{1}$$

Here, the $\delta_\mu(\cdot)$ is a non-increasing function parameterized by μ, vertex set $V = e_l^n \in M$, M is the medical events set. The event sequence is denoted by $s_n = ((e_l^n, t_l^n) : l = 1, \ldots, L_M), L_M$ is the length of s_n and $t_p^n < t_q^n$ for all $p < q$, that is to say, at time t_l^n we can observe event e_l^n in the sequence s_n.

As $\delta_\mu(\cdot)$ is a non-increasing function, so the more often and closer events i and j appear to each other in s_n, the higher W_{ij}^n is in graph G_n. We use the exceedance of the Exponential distribution $\delta_\mu(d) = \exp(-d/\mu)$ to construct the temporal graph, where d is the time interval between two events. In other words, we calculate a stronger edge weight for a smaller time interval d, when $d \le \Delta$, Otherwise, if a time interval d is larger than the threshold Δ, this event pairs will be ignored. Obviously, the weight of edge is controlled by parameters μ, Δ that can be selected according to the specific applications. That is to say, if the correlation is very small between events pairs, such as the time interval larger than 2 months, then we can set $\Delta = 2$ months, and the value of μ can be empirically set according to the average time interval between successive events.

4.2 Hybrid Collaborative Prediction Approach

We can conveniently obtain the diagnosed ICD-10 codes in time order based on the patient temporal graph representation. Thus, each patient is expressed as a vector of diagnosed diseases in time sequence. Because the diseases are not a patient choice, so the value of patient vector in medical domain is binary: a patient either has a disease (value is 1) or does not have a disease (value is 0). Using hybrid collaborative filtering algorithm that we developed to generate predictions based on a series of other similar patients with their diseases. An example of collaborative prediction model is given in Fig. 2.

Patients and diseases are represented as a matrix $\Re = I \times J$, where I refers to *all patients* and J indicts *all the possible diseases*. $J_i = (D_1, D_2, \ldots, D_z)$ represents all the diseases extracted from temporal graph of patient i and ordered by diagnosis date, as

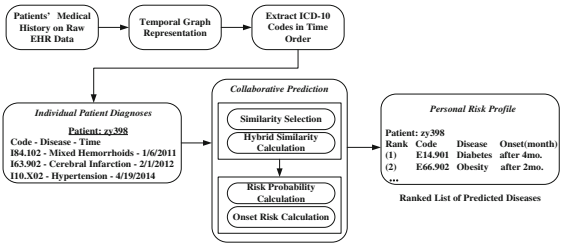

Fig. 2. Example of collaborative prediction model

shown *Patient zy398* in Fig. 2. We can see that $J = (J_1 \cup J_2 \cdots \cup J_i \mid i \in I)$. To predict future diseases for a new patient a, given $J_a = (D_1, D_2, \ldots, D_z)$ and $H_a = (D_1, D_2, \ldots, D_k)$ where $k \leq z$. The J_a is the existing diseases of patient a and H_a represent a head sequence of diseases that will be used as an input for the hybrid collaborative filtering algorithm. We define $R_a \subseteq J - H_a$ as a set of diseases to be predicted for the patient a. The goal of the collaborative prediction algorithm is to predict the probability and onset risk, then rank each disease in R_a.

Our hybrid collaborative filtering technique is derived from the similarity algorithm presented by [6], vector similarity algorithm [12] and Jaccard similarity. $S_1(a, i)$ is the first calculation measure of the similarity between testing patient a and training patients i $(i \in I)$. It is defined as the proportion of patient i's diseases to the patient a's diseases in head set:

$$S_1(a,i) = \frac{|\{D|D \in H_a \wedge D \in J_i\}|}{|H_a|}. \tag{2}$$

The second measure $S_2(a, i)$ uses vector similarity to calculate. Formally the vector similarity of patient a and i is defined in the following equation:

$$S_2(a,i) = \frac{\sum_{j \in J} v_{a,j} \cdot v_{i,j}}{\sqrt{\sum_{d \in J_a} v_{a,d}^2} \cdot \sqrt{\sum_{d \in J_i} v_{i,d}^2}}, \tag{3}$$

where $v_{a,j}$ is the value of patient a with the disease j, the possible value of v is 1 or 0.

Then, we give the third calculation method $S_3(a, i)$, which inspired by the applications of Jaccard similarity coefficient. It is suitable for all dimensions to be 0 or 1, for example, the background of this article, whether or not suffering from a certain disease. Formally the Jaccard similarity of patient a and patient i is defined in the following equation:

$$S_3(a,i) = \frac{g(|v_{a,j} = 1 \wedge v_{i,j} = 1|)}{g(|v_{a,j} = 1 \wedge v_{i,j} = 1|) + q(|v_{a,j} = 1 \wedge v_{i,j} = 0|) + r(|v_{a,j} = 0 \wedge v_{i,j} = 1|)}, \tag{4}$$

where $j \in J$, g represents the number of dimensions of patient a and i suffer from disease d, similarly, q represents the number of dimensions of only patient a have disease d, the meaning of r is opposite to q.

Thus, the ultimate similarity calculation formula is given in the following:

$$S(a,i) = \frac{1}{L} \sum_{1 \le k \le 3} S_k(a,i). \tag{5}$$

here, the value of L is 3, that is, this equation is the average of three similarity calculation measures.

For each disease d in R_a, the $N_d = \{i \mid i \in I \wedge d \in J_i\}$ represents all other patients with disease d that are similar to patient a. The probability of patient a having disease d in the future is calculated by the following equation:

$$P(a,d) = \overline{v}_d + \mu(1 - \overline{v}_d) \sum_{i \in N_d} S(a,i), \tag{6}$$

where $\overline{v}_d$ is the random expectation of disease d, i.e., $\overline{v}_d = |N_d|/|I|$, μ is a normalizing constant $\mu = 1/\sum_{i \in I} S(a,i)$. That is, the equation treats random expectation $\overline{v}_d$ as the baseline probability of each patient having disease d and adds additional risk based on similarity to other patients with disease d.

In the end, we design a formula *T(a, d)* to calculate the approximate onset time for patient a having disease d in the future, which is shown below:

$$T(a,d) = \frac{1}{|N_d|} \sum_{i \in N_d} (t_{i,d} - t_{i,x}), \tag{7}$$

here, the x is a disease that occurs d's the previous one, thus $t_{i,d}$ - $t_{i,x}$ indicts the time interval between two adjacent disease (i.e., d and x) of patient a.

4.3 Hybrid Collaborative Prediction Example

Table 1 gives an example of patient dataset in order to illustrate our hybrid collaborative prediction approach. The diseases and time interval of each patient are obtained based on the temporal graph. Thus, these diseases are ordered by the diagnosis date. For example, patient i_3 was first diagnosis as d_1, then diagnosed as d_3, the last diagnosed as d_4, and the time interval between diseases is 4 month and 3 month respectively. From Table 1, we can see the set of all patients is $I = \{i_1, i_2, i_3, i_4, i_5\}$, and the set of all possible diseases is $J = \{d_1, d_2, d_3, d_4, d_5, d_6\}$, each disease corresponds to a unique ICD-10 code. Thus, each patient can be represented a binary vector, the dimensions of this vector is six. A new patient a with diagnosed diseases d_1, d_3 and d_4 inputs the hybrid collaborative prediction measure, that is, the $H_a = \{d_1, d_3, d_4\}$, thus the target diseases to be predicted $R_a \subseteq J - H_a = \{d_2, d_5, d_6\}$.

Consider the first disease d_2 in R_a, the similar patients that have this disease d_2 are selected $N_{d2} = \{i_2, i_4, i_5\}$. We can obtain the similarity between patients (i.e., a and N_{d2}) according to the Eq. (5), so $S(a, i_2) \approx 0.26$, $S(a, i_4) \approx 0.71$, and $S(a, i_5) \approx 0.26$. Then based on the Eq. (6) to calculate the disease probability $P(a, d_2) = 0.79$. Finally, we need to compute the onset risk of disease according to the Eq. (7), $T(a, d_2) = 2$ month. Similarly, the probability of patient a developing diseases (i.e., d_5 and d_6) and onset risk corresponding to disease are as follows: d_5: $P(a, d_5) = 0.6$, $T(a, d_5) = 5$ month, d_6: $P(a, d_6) = 0.92$, $T(a, d_6) = 2$ month. Therefore the ranked list of predicted diseases for patient a is $(d6, d2, d5)$, as shown in Table 2.

Table 1. An example of patient dataset

Patient	Diagnosis
i_1	$d_4 \xrightarrow{2mo} d_6$
i_2	$d_2 \xrightarrow{3mo} d_4 \xrightarrow{2mo} d_6 \xrightarrow{4mo} d_5$
i_3	$d_1 \xrightarrow{4mo} d_3 \xrightarrow{3mo} d_4$
i_4	$d_1 \xrightarrow{2mo} d_2 \xrightarrow{1mo} d_3 \xrightarrow{5mo} d_4 \xrightarrow{3mo} d_6$
i_5	$d_3 \xrightarrow{4mo} d_2 \xrightarrow{5mo} d_5 \xrightarrow{1mo} d_6$

Table 2. Disease risk profile of new patient a

Rank	Code (ICD-10)	Disease	Onset (month)
(1)	codel	d_6	after 2 mo
(2)	code2	d_2	after 2 mo
(3)	code3	d_5	after 5 mo

5 Experiments

This section presents the experimental results to evaluate the performance of the proposed CAPM model. We apply developed approach on real world Electronic Health Records to demonstrate the improvement on predictive effectiveness.

5.1 Data Preparation

We validate the effectiveness of our presented model on a real world clinical data including the records of 92652 patients over 5 years (2011-2015). These data is collected from the medical system of a certain city in North China. The head of patient i is H_i and the head size $|H_i|$ is a parameter in our experiments. Only the patients that have $|H_i| + 1$ diseases are used for validation. In our experiments, we select patients that have at least five different diseases (i.e., $|J_i| \geq 5$) in our database so that there could be sufficient medical history for both training and evaluation. As a result, 7372 patients are selected in the final who meet the condition.

This paper use the inpatient diagnosis information of International Classification of Disease 10 (ICD-10) codes and the medication information according to drug action and corresponding timestamps to construct the temporal sequences. Then the patient temporal graphs are constructed from those sequences in terms of *Definition* 1. Figure 3 gives an example of one patient's temporal graph, which contains five diseases that are uniquely encoded by ICD-10. Then, we extract these codes in time order based on temporal graph representation, which is used to be the input of the hybrid collaborative prediction approach.

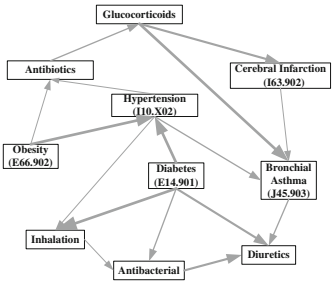

Fig. 3. Example of one patient's temporal graph

To evaluate the hybrid collaborative prediction measure, we use a leave-one-patient-out validation strategy similar to [11]. One active patient i is taken out and the other patients are used for training every time. Then the $|H_i|$ diseases of patients i are fed into the trained hybrid collaborative prediction model. The remaining $|J_i|$-$|H_i|$ diseases are considered as future diseases and used for evaluation. The top-K diseases in the ranked list of predicted conditions are considered. Above process is repeated for each patient.

5.2 Validation of the Developed Approach

Baselines. In order to evaluate our model introduced in Sect. 4, the following baseline methods for comparison purpose will be considered:

(1) *Baseline-1 (BL_1)*, this method implements the similarity calculation algorithm proposed in [12], which only uses vector similarity (i.e., Eq. (3)) to calculate the similarity among patients.
(2) *Baseline-2 (BL_2)*, this way implements the similarity calculation approach proposed in [6], which only uses Eq. (2) to compute the similarity between patients.
(3) *Baseline-3 (BL_3)*, this method only uses the Eq. (4) that we designed to calculate the similarity of different patients.

Metrics. We use *Coverage* and *Average rank* to assess the prediction performance for each patient. The content of the two metrics are as follows statement:

(1) *Coverage*. It is defined as the percentage of diseases for which a prediction is ranked. That is to say, coverage is the proportion of correct future diseases in the top-K ranked list to the total number of correct future diseases, as shown in the Eq. (8). Apparently, we desire to capture as many future diseases as possible, so the higher coverage, the better prediction performance of approach.

$$Coverage = \frac{\#Predicted\ target\ diseases}{\#Total\ target\ diseases}. \tag{8}$$

(2) *Average rank*. It is satisfactory for future diseases to have the low rank positions. Thus, we use the average rank of all correct future diseases in the ranked list for this patient as an evaluation metric. As shown in the Eq. (9). Ideally, if a patient actually has the diseases, which should be near the top of ranked list, so that they are most probably to be noticed and used.

$$Average\ rank = \frac{\#Total\ target\ number}{\#Total\ target\ diseases}. \tag{9}$$

Results. Table 3 displays the prediction performance of CAPM compared with the baselines ranking, where the head size is 3. Results on the top 20 and top 100 ranks are more significant, because the medical experts or other users are impossible to consider a large portion of the list. The hybrid collaborative assessment prediction model achieves a coverage value of 49% and 76% for top-20 and top-100 ranked lists respectively. From the top-20 we can observe that our developed approach CAPM significantly improve the predictive performance compared to the baseline BL_3 method, the coverage obtained by BL_3 is only 32% (a gain of 8%). Similarly, from top-100, we also can observe that the CAPM outperforms other three methods, compared to the basic BL_1 method that achieves 60% (a gain of 16%). Table 4 shows the specific examples of predictions using our developed model. Because this paper sets the parameter of $|H_i| = 3$, the number of Diagnosed Diseases is three in the Table 4. As a summary, the experimental results have demonstrated the effectiveness of our developed model on a real EHR data, which can achieve better prediction performance compared to the baseline methods.

Table 3. Prediction performance of CAPM compared with the baseline ranking

Comparison of methods				
	BL_1	BL_2	BL_3	CAPM
Top 20				
Coverage	43%	47%	41%	49%
Average rank	7.81	7.22	6.80	5.76
Top 100				
Coverage	60%	70%	68%	76%
Average rank	26.63	21.32	22.04	20.19
All				
Coverage	94%	96%	89%	99%
Average rank	170.39	122.19	91.19	90.37

Table 4. Example of future predictions for individual patients

Patient ID	Diagnosed diseases	Top 2 predicted diseases
zy398	Mixed Hemorrhoids(I84.102), Cerebral Infarction (I63.902), Hypertension(I10.X02)	Diabetes(E14.901), Obesity(E66.902)
zyl5177	Esophagus Cancer(Z98.850), Liver Cancer (C22.902), Hypertension(I10.X02)	Pneumonia(J12.901), Lung Cancer(C34.904)
zy11138	Heart Disease(I11.901), Obesity(E66.902), Esophagitis(K22.103)	Diabetes(E14.901), Hypertension(I10.X02)

6 Conclusion

This paper has proposed a Collaborative Assessment Prediction Model (CAPM) based on patient temporal graph to predict future disease risks. The CAMP provided a unified temporal graph representation by summarizing each patient's longitudinal raw EHRs, which is informative for a variety of challenging analytic tasks because it can capture temporal relationships between clinical events. Moreover, this paper developed a hybrid collaborative prediction approach to calculate the similarity among patients, which only use ICD-10 codes extracted from patient temporal graph. In addition, we have proposed an approach to calculate the onset risk of each predicted disease. The patient's disease risk profile obtained by our model not only includes the ranked list of diseases, but also contains the onset risk of each disease. The experimental results have shown that the proposed model could improve the effectiveness compared to the basic prediction methods in our real world EHR dataset.

Acknowledgement. This work is partially supported by NSFC No. 61303005, 61572295; the Innovation Method Fund of China No. 2015IM010200; SDNSFC No. ZR2014FM031; the Science and Technology Development Plan Project of Shandong Province No. 2014GGX101019, 2015GGX101007, 2015GGX 101015; the Shandong Province Independent Innovation Major Special Project No. 2015ZDJQ010 02, 2015ZDXX0201B03; the Fundamental Research Funds of Shandong University No. 2014JC025, 2015JC031.

References

1. Laura, B.M.: Data-Driven Healthcare: How Analytics and BI are Transforming the Industry. Wiley (2014)
2. Gotz, D., Wang, F., Perer, A.: A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. Biomed. Inform. **48**, 148–159 (2014)
3. Davis, D.A., Chawla, N.V.: Predicting individual disease risk based on medical history. In: Information and Knowledge Management, pp. 769–778 (2008)
4. Dentino, B., Davis, D., Chawla, N.V.: HealthCareND: leveraging EHR and ARE for prospective healthcare. In: Health Informatics Symposium, pp. 841–844 (2010)

5. Liu, C., Zhang, K., Xiong, H., Jiang, G., Yang, Q.: Temporal skeletonization on sequential data: patterns, categorization, and visualization. In: KDD, pp. 211–223 (2014)
6. Ji, X., Chun, S.A., Geller, Z., Oria, V.: Collaborative and trajectory prediction models of medical conditions by mining patients' Social Data. In: BIBM, pp. 695–700 (2015)
7. Zhou, J.Y., Wang, F., Hu, J.Y., Ye, J.P.: From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records. In: SIGKDD, pp. 135–144 (2014)
8. Ooi, B.C., Tan, K.-L., Tran, Q. T., Yip, J.W.L., Chen, G., Ling,Z.J., Nguyen, T., Tung, A.K. H., Zhang, M.: Contextual crowd intelligence. In: SIGKDD, pp. 39–46 (2014)
9. Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Comput. **7**, 76–80 (2003)
10. Hofmann, T.: Latent semantic models for collaborative filtering. Trans. Inf. Syst. **22**, 89–115 (2003)
11. Xia, P., Liu, B., Sun, Y., Chen, C.: Reciprocal recommendation system for online dating. Soc. Netw. Anal. Mining. **9**, 234–241 (2015)
12. Davis, D.A., Chawla, N.V., Christakis, N.A., Barabási, A.L.: Time to CARE: a collaborative engine for practical disease prediction. Data Min. Knowl. Disc. **20**, 388–415 (2010)
13. Sun, J., Wang, F., Hu, J., Edabollahi, S.: Supervised patient similarity measure of heterogeneous patient records. In: SIGKDD, pp. 16–24 (2012)
14. Hussein, A.S., Omar, W.M., Li, X., Hatem, M.A.: Smart collaboration framework for managing chronic disease using recommender system. Health Syst. **3**, 12–17 (2014)