

Directed disease networks to facilitate multiple-disease risk assessment modeling

Tingyan Wang^{a,b}, Robin G. Qiu^{c,*}, Ming Yu^a, Runtong Zhang^d

^a Health Care Services Research Center, Department of Industrial Engineering, Tsinghua University, Beijing 100084, China

^b Nuffield Department of Medicine, University of Oxford, South Parks Road, Oxford OX1 3SY, United Kingdom

^c Big Data Lab, Division of Engineering and Information Science, The Pennsylvania State University, Great Valley, Malvern, PA 19355, United States

^d Department of Information Management, School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China

ARTICLE INFO

Keywords:

Multiple-disease risk prediction
Health risk assessment
Disability adjusted life year
Directed network
Disease temporal relations

ABSTRACT

We investigate multiple disease risk prediction modeling, aimed at assessing future disease risks for an individual who is ready for discharge after hospitalization. We propose a novel framework that combines directed disease network and recommendation system techniques to substantially enhance multiple disease risk predictive modeling. Firstly, a directed disease network considering temporal information is developed. Then based on this directed disease network, we look into different disease risk score computing approaches. We validate the proposed approaches with two real-world datasets from two independent hospitals. The predicted results can be promisingly utilized as a reference for medical experts to offer effective healthcare guidance for both inpatients and outpatients. The proposed framework can also be utilized for developing an innovative tool that helps individuals create and maintain a better healthcare plan over time.

1. Introduction

Identification of disease risks and intervention at the earliest stage can lead to better medical results and lower medical cost [1]. Knowing disease risks in time helps not only improve healthcare services but also reduce re-admission rate [2], which facilitates building strong and healthy communities over time. However, currently millions of individuals suffer from late-diagnosed chronic diseases [3], which causes heavy burdens to the society [4]. Recently, the assessment of health risks has drawn much attention in academia and practice, yet the accuracy of health risk evaluation remains one of the main challenges in healthcare research [1]. Therefore, to improve the accuracy of health risk assessment for an individual, it is essential to study multi-disease risk prediction modeling in a systematic way.

As hospital information systems are widely adopted, there is a tremendous amount of health-related information in electronic health records (EHRs), which could be well utilized to benefit patient-centered healthcare. For example, medical records containing information concerning disease correlations and progression are particularly fundamental to multiple disease risk evaluations [5–9]. Many researchers have exploited comorbidity relationships among diseases based on medical histories of patients for multiple disease risk prediction modeling [10–14], which focused on disease prediction modeling through

investigating concurrence of diseases while excluding their temporal relations among afflicted diseases. Researchers recently started to take into account temporal information to study directed disease networks [6,7] or directly incorporate disease temporal relations into disease risk predictive modeling [4,15,16]. However, there is limited literature studying multiple disease prediction modeling using directed disease networks to fully mine and leverage the temporal information in EHRs.

In this study, we formulate the multiple disease prediction as a medical recommendation system problem and propose a novel framework that combines directed disease network and recommendation system techniques for multiple disease risk prediction modeling. To validate the applicability of the framework, we apply the proposed framework to two archived medical datasets provided by two general hospitals in Beijing and Shenzhen, China, respectively.

The remaining paper is organized as follows. Section 2 presents the problem under study. Section 3 reviews the related literature. Section 4 then details the proposed multiple disease risk modeling. Data analysis and results based on hospital datasets are presented in Section 5. Lastly, Section 6 discusses the application of the proposed framework, and the contributions and limitations of this study, and Section 7 concludes this study.

* Corresponding author.

E-mail address: robinqiu@psu.edu (R.G. Qiu).

<https://doi.org/10.1016/j.dss.2019.113171>

Received 16 April 2019; Received in revised form 22 October 2019; Accepted 22 October 2019

Available online 23 October 2019

0167-9236/ © 2019 Elsevier B.V. All rights reserved.

Table 1

A sample of medical records of a patient.

Patient ID	n^{th} visit	Diagnosis codes
P1	1	N18.001, I10, I15.101, N03.801, J92.901, J98.401, E83.501, D64.802, N28.101
P1	2	N18.001, I10, I15.101, N03.801, J92.901, D64.802
P1	3	N18.001, I10, I15.101, N03.801, E04.101, D64.802

2. Problem statement

In general, there are two different ways of conducting disease prediction modeling: (a) single disease prediction modeling, which involves only one type of disease in risk assessment, (b) multiple disease prediction modeling, which aims to identify all diseases that an individual might develop. The former is a disease-centric model that predicts the likelihood of an individual getting a specific disease regardless of whether other diseases will occur to this individual. The latter is a patient-centric model that focuses on assessing multiple disease risks for an individual simultaneously. In this study, we concentrate on multiple disease prediction modeling. It is worth mentioning that some studies involve several diseases. However, their prediction approaches are either based on one disease at a time [17] or a limited number of specific diseases simultaneously [18,19]. Because this study focuses on simultaneously assessing all the future disease risks for a patient, their studies are different from our work.

Many patients experienced hospital re-admissions due to suboptimal healthcare, chronic conditions, or suffering additional diseases [2,20]. Each hospital visit of a patient produces a set of medical diagnoses. As shown in Table 1, each row represents a hospital visit of a patient. The first column indicates the patient's identification, the second column denotes the sequence number of each hospital visit, and the third column includes the patient's diagnosis in each hospital visit. Patients' diagnoses in the raw datasets under study have been encoded according to the 10th Revision of International Statistical Classification of Diseases and Related Health Problems (ICD-10) [21,22].

In this study, the question is how to predict disease risks of a patient over time based on the patient's medical history. This question is motivated by the fact that doctors are interested in knowing which diseases a patient might suffer in the future so that better medical advice can be provided to the patient when she/he gets discharged from the hospital. In this study, we formulate this prediction problem as a recommendation system rather than a multi-label classification problem. The input to our model is medical records of patients, and the output is a list of ordered diseases that the target patient may be afflicted with on the subsequent hospital visit. As an excessively long recommended list of diseases would be inappropriate in practice, we provide a recommended list with a fixed length, denoted as Q that can be changed, depending on the needs of medical experts, professionals, or the patients.

3. Related work

Lots of researchers have studied single disease prediction modeling, e.g., heart disease prediction [23]. The methods used for a single disease prediction problem are classifiers, such as neural networks, decision tree, Naïve Bayes, random forest, logistic regression, support vector machine, Bayesian network, deep learning and so on [17,24–26]. Two metrics, relative risk and odds ratio, are frequently used in single disease risk assessment to quantify the relationship between an independent variable and the dependent variable of interest, i.e., the causal effect or exposure effect for a single disease [27,28].

To clearly show how we enrich the literature in addressing multiple disease prediction problems, we summarize the state-of-the-art in the field in Table 2. It shows that there are limited studies using directed

Table 2
Studies for multiple disease prediction.

	Recommendation algorithms	Undirected disease network	Directed disease network	Other methods
Without incorporating temporal information	Davis et al. [10]; Chawla and Davis [2]; Dasgupta and Chawla [29]	Steinhaeuser and Chawla [11]; Folino et al. [34]; Folino and Pizzuti [13]; Lakshmi and Vadivu [9]	–	Folino and Pizzuti [12]; McCormick et al. [35]; Rider and Chawla [14]; Chang et al. [18]; Wang et al. [36]; Li et al. [37]; Bayati et al. [1]; Maxwell et al. [19]
Incorporating temporal information	Davis et al. [30]; Ji et al. [31–33]; Nasiri et al. [4]	–	Our study; Jensen et al. [6]; Kannan et al. [7]	Folino and Pizzuti [16,17]; Miotto et al. [39]; Choi et al. [40]; Razavian et al. [41]; Kim et al. [42]; Ma et al. [43]; Nguyen et al. [44]

Note: the studies are listed in the order of their published time.

disease networks to leverage temporal information in multi-disease prediction modeling. Next, the studies in Table 2 are reviewed in detail.

Future disease risk assessments for an individual using recommendation algorithms while without considering disease temporal information have been exploited. Davis et al. [10] proposed a model called Collaborative Assessment and Recommendation Engine (CARE), which is considered as one of the earliest studies that applied collaborative filtering to predict multi-disease risks. They identified top disease risks for a patient by comparing the patient's profile with similar patients' profiles. They also developed an iterative version to enhance CARE, called ICARE. The results demonstrate that collaborative filtering has practical potential in multiple disease prediction. Chawla and Davis [2] promoted the application of CARE and ICARE for personalized healthcare. By incorporating medication information, Dasgupta and Chawla [29] developed medCARE and combinedCARE based on CARE and articulated that combinedCARE improved outcomes.

By contrast, with the support of recommendation algorithms other researchers have been trying to directly incorporate temporal information among diseases for multiple disease prediction. Davis et al. [30] extended ICARE to time-sensitive ICARE by leveraging longitudinal EHRs data. Ji et al. [31–33] proposed a framework using collaborative filtering that not only predicts medical condition incidences but also reveals disease progression trajectories. Nasiri et al. [4] designed a recommendation algorithm based on tensor factorization to predict disease risks for patients. By incorporating the time dimension, this algorithm shows better results compared with ICARE.

To take advantage of the phenotypic information of diseases, some researchers have tried to develop a multi-disease risk prediction model through construction of disease networks using patients' medical histories. Steinhäuser and Chawla [11] built an undirected disease network, in which nodes are the diseases and each edge weight is assigned with the ratio of co-occurrence of two diseases by the same patient. Then nearest neighbor method and depth-first search technique were applied to assess disease risks for a patient. Similarly, Folino et al. [34] designed an undirected phenotypic disease network, in which each edge is labeled with the number of patients who share the same diseases. Then they applied association rule analysis to find disease frequent patterns. Folino and Pizzuti [13] also constructed a comorbidity network of diseases and applied link prediction approaches to inferring new disease correlations among their study cohort. Lakshmi and Vadivu [9] developed a method based on weighted association rule mining to predict disease comorbidities using both clinical and molecular data. Although the undirected network-based methods can incorporate the comorbidity relations of diseases, unfortunately, models based on undirected networks can't integrate the temporal information between diseases.

In addition to recommendation algorithms and undirected disease network, there are some other methods without incorporating temporal information, such as association rule analysis [12,33], topic model [14], classifiers [1,18,36,37] and deep learning [39], proposed for the multiple disease prediction. Folino and Pizzuti [12] developed a system called CORE (COMorbidity-based Recommendation Engine), by clustering patients' medical records first and then applying association rule analysis in the disease risk prediction modeling. Rider and Chawla [14] proposed a modified Dirichlet process mixture model to allow aggregating distinct EHR datasets so as to enhance the performance of disease risk predictions; McCormick et al. [35] proposed a hierarchical association rule model to generate association rules for predicting future medical conditions; Maxwell et al. [19] used deep neural networks to predict 8 types of disease risks. However, models based on classifiers are generally like 'black boxes', which lack interpretability [3,38].

To incorporate the temporal information, Folino and Pizzuti [15,16] proposed CORE⁺ by integrating Markov models and sequential pattern mining instead of frequent pattern mining into CORE. With combining more models and considering the hospital visit sequence of a patient, CORE⁺ is superior to CORE with respect to the prediction accuracy.

Recently, scholars have begun to leverage deep learning classification algorithms to predict multi-disease risks or a specific set of diseases: Miotto et al. [39] used unsupervised deep learning to derive a patient representation from various types of medical records in EHRs for predicting disease risks; Choi et al. [40] applied gated recurrent units based recurrent neural network (RNN) to predict the diagnoses and medications of the subsequent visit for a patient; Razavian et al. [41] used neural networks to predict disease onsets based on longitudinal measurements of lab tests; Kim et al. [42] applied deep attention networks to predict vascular diseases based on diagnosis and pharmacy codes; Ma et al. [43] employed bidirectional RNNs to predict diagnosis codes; Nguyen et al. [44] modeled the sequences of patients' visits using RNN to predict disease risks for patients with diabetes and mental health problems. The achievements from these studies based on RNN were remarkable to some extent. However, RNN-based approaches may not fully leverage all the previous visit information of patients, which requires further investigations [43].

Recently, some researchers have explored clinical relationships and disease trajectories by constructing directed networks: Jensen et al. [6] derived disease trajectories by investigating temporal patterns among pairs of diagnoses; Kannan et al. [7] introduced a causal information fraction measure to determine the directionality of disease trajectories. In this study, rather than only focusing on disease trajectories, we use a directed disease network to facilitate multiple disease prediction modeling. With combining directed disease network and recommendation system techniques, we can fully incorporate and leverage chronological orders between patients' successive hospital visits.

4. Methodology

In this section, we introduce our multi-disease prediction modeling based on our preliminary study [45]. Fig. 1 depicts the modeling flow and main activities in the proposed modeling methodology. In the following sections, we describe each step of the proposed methodology in detail.

4.1. Grouping and ranking diseases with disability-adjusted life year

Let's firstly explain the reason for grouping and ranking diseases. ICD-10 is a standard dictionary for the classification of diseases and related health problems. Each disease has a code in the dictionary. The top categories of ICD-10 codes and the code descriptions are available online [22].

According to ICD-10, there are over ten thousand diseases marked with four characters (excluding the decimal point). However, from the perspective of health risk management, there is no need for us to conduct disease risk analysis at this excessively detailed level. In other words, diseases can be systematically categorized into levels for easy interpretation and management in hospitals.

Since severity levels of disease risks vary with countries or geographical regions, they should be identified within the group or population under study. Therefore, country-based severity levels can help identify what diseases affect patients more severely and how those diseases can be geographically controlled and managed in an effective manner. In light of this, we aggregate diseases into groups using Disability-Adjusted Life Year (DALY), which is a comprehensive metric of disease burdens designed by WHO [46]. DALY has been widely promoted by WHO to quantify the burden of disease from mortality and morbidity. As defined by WHO, "DALY for a disease or health condition are calculated as the sum of the Years of Life Lost due to premature mortality in the population and the Years Lost due to Disability for people living with the health condition or its consequences". One DALY can be thought of as one lost year of "healthy" life due to the affliction of a disease. Hence, DALY is an excellent reference used to assess the burden or severity of a disease.

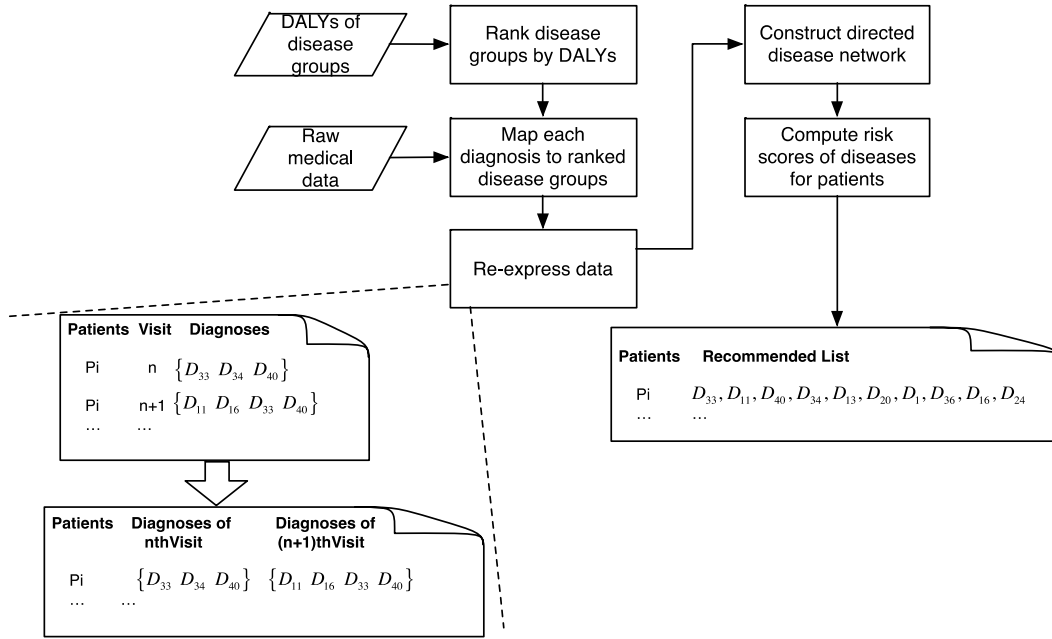


Fig. 1. The process overview of multiple disease prediction modeling in this study.

DALY has been applied to estimating disease burdens. Particularly, DALY has been used to classify diseases into categories easily manageable in hospitals. WHO Global Health Estimates (GHE) provides datasets of DALY estimates for all WHO members [47]. Hence, we adopt DALY in our multiple disease risk prediction modeling. Essentially, each category of DALY estimates represents a group of diseases, i.e., a set of ICD codes. There are 125 disease groups sorted by the DALY estimates for China population, which can be represented in a descending order, such as D_1, D_2, \dots, D_{125} . These ranked disease groups will then be used as nodes to construct a directed network. Apparently, the proposed methods can be easily extended to other countries or globally.

4.2. Data preprocessing methods

Two steps are used for preprocessing a raw dataset before a model can be developed:

- (1) Mapping patients' diagnostic results to the ranked disease groups of DALY

As for raw medical records, the ICD-10 codes of medical diagnoses have more than 4 characters (excluding the decimal point) in length, covering over ten thousand diseases. To use the ranked disease groups of DALY as the nodes of a directed network, these medical diagnostic results need to be aggregated and mapped into the above-mentioned ranked disease groups. As a result, each patient has a set of diseases with ranked numbers in a chronological order.

- (2) Data transformation for directed disease network design

To construct a directed disease network, we must transform the data in a way of showing the temporal relations between a patient's two consecutive hospital visits. As illustrated in Fig. 2, the column named *id* denotes the hospital identification of a patient; the column named n^{th} indicates the n^{th} visit of the patient's hospital visits. For the column D_i , $i = 1, 2, \dots, 125$, the value is 1 if a patient suffers from disease D_i , and 0 otherwise.

In Fig. 2, for matrix U , we merge the records of the n^{th} and the $(n+1)^{th}$ hospital visit into one record, thus we obtain matrix V . For

example, as illustrated in Fig. 2, the records of the first visit and the second visit of the first patient can be merged into one record. Therefore, with the data transformation by converting two records of consecutive visits into one, matrix V can clearly describe the temporal relations of diseases between a patient's two consecutive visits.

4.3. Directed disease network design


In this section, we illustrate how to develop a directed network by leveraging chronological orders between successive hospital visits to investigate temporal relations among diseases.

A directed disease network can be easily developed. A vertex x in the network denotes a disease. A directed arc (x, y) indicates disease x pointing to disease y , which is constructed from patients' two consecutive visits. The flow of the directed arc (x, y) , called the outflow of vertex x to vertex y , shows that the number of patients with disease x for a hospital visit and then with disease y diagnosed in their subsequent visits. The total outflows of vertex x are equal to the flows coming out from this vertex to other vertices.

To make it easy to understand, we use a simplified network to explain the developed disease network under study. As displayed in Fig. 3(a), there are three vertices, which represent stroke, liver cancer, and diabetes mellitus. The numbers beside directed arcs indicate flows among various vertices. For example, the outflows from stroke to liver cancer are 11, which means that there are 11 patients with stroke in their last visits and liver cancer in their current visits. The total number of outflows of stroke is 2201, which includes flows to other vertices that are not shown in the simplified network. It is worth mentioning that in this study directed arcs in the network are not causal links, i.e., the flows among disease nodes simply stand for the correlations from their temporal perspectives rather than their causal relationships. Fig. 3(b) shows the network of top 20 disease groups under study. Particularly, the thickness of each edge denotes the intensity of an outflow or an inflow. Obviously, it gets more complicated and difficult to read if we draw all the disease groups to build a complete network. Therefore, the figure for the network with all disease groups is omitted here.

With the complexity of the temporal relations among diseases, the question now is how to assess disease risks that an individual will have on his/her next hospital visit.

id	n^{th}	D_1	D_2	D_3	\dots	D_{125}
1	1	0	1	1	\dots	0
1	2	1	1	0	\dots	1
1	3	1	1	0	\dots	0
2	1	0	1	0	\dots	1
2	2	1	0	1	\dots	0
\vdots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots



id	n^{th}	D_1	D_2	D_3	\dots	D_{125}	id	n^{th}	D_1	D_2	D_3	\dots	D_{125}
1	1	0	1	1	\dots	0	1	2	1	1	0	\dots	1
1	2	1	1	0	\dots	1	1	3	1	1	0	\dots	0
2	1	0	1	0	\dots	1	2	2	1	0	1	\dots	0
\vdots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots

Fig. 2. Transforming data to construct a directed disease network.

4.4. Disease risk score computing for disease risk predictions

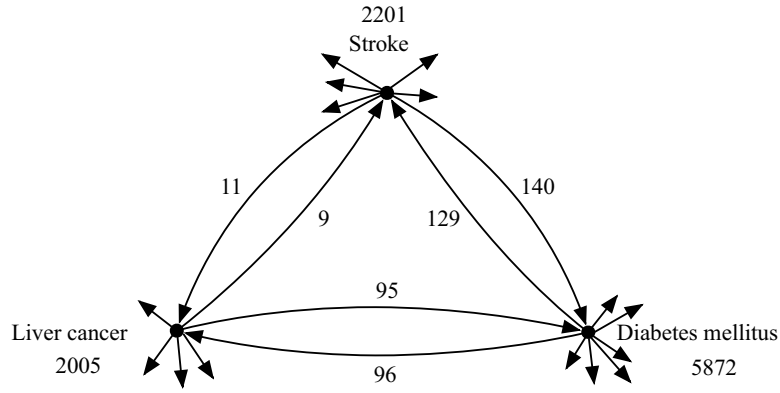
Based on the disease relations developed in the directed network and the medical history of patient i , two scoring approaches are first investigated to calculate the likelihoods of developing different diseases by the target patient i . Then an ordered list of diseases, which are sorted by the risk scores, can be predicted for the target patient i .

4.4.1. Approach 1: disease temporal link method

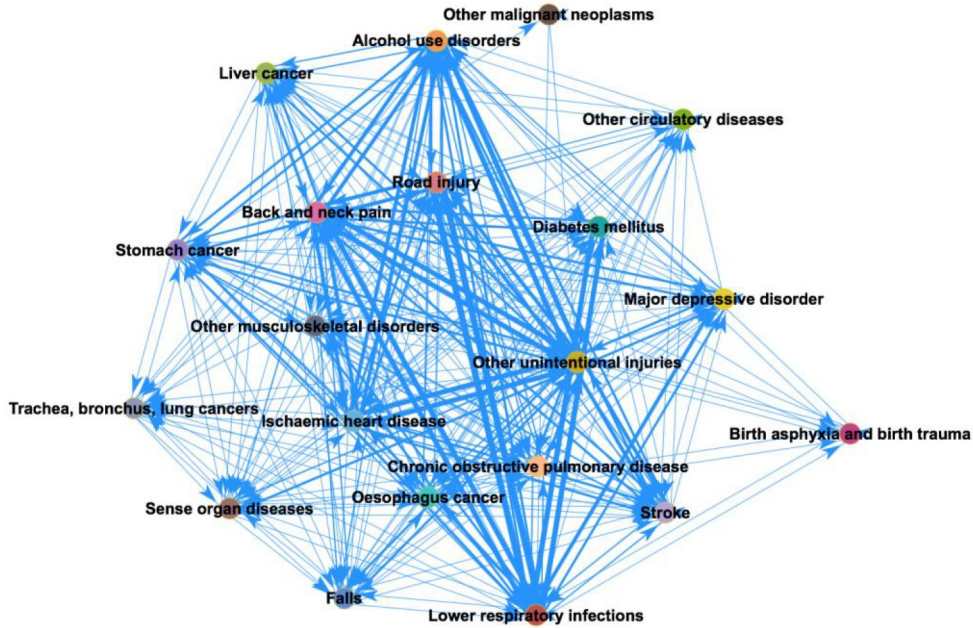
The first approach directly leverages the temporal relations between

nodes in the network, which is denoted as Approach 1 or disease temporal link method (DTLM). Based on the diseases in a patient's medical history, we can compute the likelihoods of the next developing diseases of the patient. Considering patient i with M diseases in the medical history, denoted as $H = \{h_1, h_2, \dots, h_M\}$, the basic idea of Approach 1 is that the risk score of the target patient i developing a disease k is computed according to the relations between disease k and each disease in the patient's medical history, i.e.,

$$Score_k^{(i)} = g(n_{m \rightarrow k}) \quad (1)$$



(a) A simplified view of three diseases with outflows



(b) The network of top 20 diseases in the dataset under study

Fig. 3. Sub-graph of the directed network in the dataset under study.

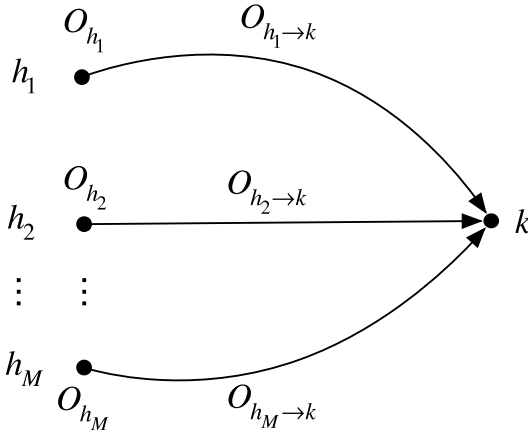


Fig. 4. illustration of a patient's disease temporal link.

where $h_m \in H$, $r_{h_m \rightarrow k}$ indicates a defined temporal relation between disease h_m and disease k , and $g(\cdot)$ represents a function of disease temporal relations.

In Approach 1, we use the flows between a disease in a patient's medical history and other diseases to define the temporal relations between nodes. Fig. 4 shows a directed sub-graph with outflows on vertices and arcs. The total outflows of the vertices h_1, h_2, \dots, h_M are $O_{h_1}, O_{h_2}, \dots, O_{h_M}$ respectively, and their outflows to vertex k are $O_{h_1 \rightarrow k}, O_{h_2 \rightarrow k}, \dots, O_{h_M \rightarrow k}$. The disease temporal link approach is that we compute the risk score of vertex k using the outflow of each node in the set H directing to vertex k . Accordingly, Eq. (1) can be rewritten as the average risk score of patient i developing disease k on the next hospital visit, which is calculated as follows:

$$Score_k^{(i)} = \frac{1}{M} \sum_{m=1}^M \frac{O_{h_m \rightarrow k}}{O_{h_m}} \quad (2)$$

where $h_m \in H$, O_{h_m} denotes the total outflows of vertex h_m , $m = 1, \dots, M$, and $O_{h_m \rightarrow k}$ indicates the outflow from disease h_m to disease k .

4.4.2. Approach 2: aggregated disease temporal links method

Rather than only incorporating the temporal relations between single nodes, we consider the temporal relations between each subset of the target patient's medical history and disease k , and then derive the risk score of developing disease k by aggregating risk scores from all the subsets. We denoted this approach as Approach 2 or aggregated disease temporal links method (ADTLM).

Like in Approach 1, we consider patient i with M diseases in the medical history, denoted as $H = \{h_1, h_2, \dots, h_M\}$. In Approach 2, we first figure out all the subsets of the target patient's medical history and then calculate the score for each subset. Finally, we aggregate the scores of these subsets with different weights and obtain the risk score of the target patient developing disease k :

$$Score_k^{(i)} = f(w_{H_{sub}} \cdot SubScore_{H_{sub} \rightarrow k}^{(i)}) \quad (3)$$

where $H_{sub} \subseteq H$, $H_{sub} \rightarrow k$ indicates a defined temporal relation between the subset H_{sub} and disease k , $w_{H_{sub}}$ are weights derived according to the sizes of subsets of the target patient's medical history, $SubScore_{H_{sub} \rightarrow k}^{(i)}$ indicates the score derived from each subset and $f(\cdot)$ represents a function that combines scores from different subsets.

A subset is defined as a set of the same disease(s) suffered by the target patient and other patients. The size of a subset indicates the number of diseases that other patients share with the target patient in their medical histories. In this study, we define a similarity metric only based on patients' medical histories, called "medical history similarity", which is different from the similarity based on demographic information that is commonly used in a recommendation system. The metric "medical history similarity" between two patients is quantified as the

set size of the shared diseases in these two patients' medical histories. Obviously, the more diseases shared by two patients in the medical histories, the more similar their medical histories are. As mentioned before, the basic idea of Approach 2 is to leverage the temporal relations between subsets of the target patient's medical history and disease k to derive the risk score of the target patient developing disease k . Therefore, the larger the medical history similarity between the target patient and another patient is, the greater the likelihood that the target patient will develop an identical disease with this patient. Correspondingly, we assign different weights to subsets with various sizes for Approach 2 according to the medical history similarity.

To assign different weights to subsets with different sizes, it is necessary to group medical history subsets with the same size into an identical category. Given patient i with the medical history $H = \{h_1, h_2, \dots, h_M\}$, we get the subsets of the medical history, such as $\{h_1\}, \{h_2\}, \dots, \{h_1, h_2\}, \{h_1, h_3\}, \dots, \{h_1, h_2, h_3\}, \{h_1, h_2, h_4\}, \dots, \{h_1, h_2, \dots, h_M\}$. Then these medical history subsets are classified into multiple categories J_1, J_2, \dots, J_M according to their sizes. For example, J_1 is a category of subsets, which has only one element. By the same token, $J_m, m = 1, 2, \dots, M$, is a category of subsets, which has m diseases. Mathematically, we have

$$J_m = \{A \subseteq H \mid Card(A) = m\} \quad (4)$$

where $m = 1, 2, \dots, M$, $Card(A)$ is the number of elements in set A .

In Approach 2, we use the flows between subsets in a patient's medical history and disease k to define temporal relations (Fig. 5). For each subset $sub_j \in J_m, j = 1, 2, \dots, Card(J_m)$, we figure out the total outflow, and their outflows directing to vertex k . Thus, the average score of each category is:

$$SubScore_{J_m \rightarrow k}^{(i)} = \frac{1}{Card(J_m)} \sum_{j=1}^{Card(J_m)} \psi(sub_j \rightarrow k) \quad (5)$$

$$\psi(sub_j \rightarrow k) = \begin{cases} \frac{O_{sub_j \rightarrow k}}{O_{sub_j}}, & O_{sub_j} \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where $sub_j \in J_m$, O_{sub_j} indicates the total outflow of the subset sub_j , $O_{sub_j \rightarrow k}$ denotes the outflow from the subset sub_j to disease k , and $Card(J_m)$ is the number of subsets in category $J_m, m = 1, 2, \dots, M$.

As we have classified medical history subsets with an identical number of elements into the same category, medical history subsets in the same category are then assigned with the same weights. In this study, the weights are calculated based on an exponential function of the sizes of subsets. In detail, given m denotes the size of a medical history subset, i.e., the number of the diseases in the medical history that another patient sharing with the target patient i , then the weight of the category is e^m . Correspondingly, the normalized weight of each category is

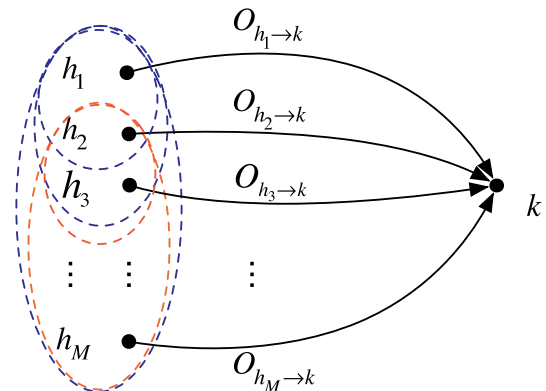


Fig. 5. illustration of a patient's aggregated disease temporal links.

$$w_{J_m} = \frac{e^m}{\sum_{m=1}^M e^m} \quad (7)$$

where $m = 1, 2, \dots, M$. Thus, the risk score of patient i with disease k in the next hospital visit is:

$$Score_k^{(i)} = \sum_{m=1}^M (w_{J_m} \cdot SubScore_{J_m \rightarrow k}^{(i)}) \quad (8)$$

Based on Eqs. (5), (6), and (7), Eq. (8) can be further defined as:

$$Score_k^{(i)} = \frac{1}{\sum_{m=1}^M e^m} \cdot \sum_{m=1}^M \left[\frac{e^m}{Card(J_m)} \cdot \sum_{j=1}^{Card(J_m)} \psi(sub_j \rightarrow k) \right] \quad (8-1)$$

$$\psi(sub_j \rightarrow k) = \begin{cases} \frac{O_{sub_j \rightarrow k}}{O_{sub_j}}, & O_{sub_j} \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (8-2)$$

In summary, in Approach 1 (DTLM) we simply use the links among single nodes to derive risk scores. By contrast, in Approach 2 (ADTLM) we consider a more complex way based on the relationship of disease sets and disease nodes to calculate risk scores, which fully leverages the temporal disease relation information of other similar patients to assess disease risks for the target patient.

4.4.3. Predicted diseases list generation

Through the above-discussed risk score computing schemes, we can generate a set of disease scores for patient i , i.e., $\{Score_1^{(i)}, Score_2^{(i)}, \dots, Score_K^{(i)}\}$. Let $\langle Score_{(1)}^{(i)}, Score_{(2)}^{(i)}, \dots, Score_{(K)}^{(i)} \rangle$ denote the sequence of the scores $Score_k^{(i)}$, $k = 1, 2, \dots, K$, then we can obtain a corresponding sequence of diseases, denoted as $\varphi^{(i)} = \langle \varphi_1^{(i)}, \varphi_2^{(i)}, \dots, \varphi_K^{(i)} \rangle$. As mentioned in the problem statement, we provide a predicted list with a fixed length of Q . In other words, only top Q diseases in the sequence $\varphi^{(i)} = \langle \varphi_1^{(i)}, \varphi_2^{(i)}, \dots, \varphi_K^{(i)} \rangle$ of patient i will be finally recommended to healthcare professionals.

4.5. Evaluation metrics

The result of our prediction model is an ordered list of disease risks for a patient. It needs an assessment method to evaluate the accuracy of each recommended list with respect to the corresponding true values in the dataset. We adopt an accuracy evaluation approach based on the half-life accuracy evaluation [48]. The basic idea of the half-life accuracy method is to assign different weights to various positions on the recommended list.

A set of diseases that patient i will have when visiting the hospital next time, is called a true value set of patient i , denoted as $\theta^{(i)} = \{\theta_1^{(i)}, \theta_2^{(i)}, \dots, \theta_G^{(i)}\}$. The whole predicted list can be denoted as $\varphi^{(i)} = \langle \varphi_1^{(i)}, \varphi_2^{(i)}, \dots, \varphi_K^{(i)} \rangle$, while as mentioned earlier only top Q diseases on the list will be recommended. The value of a position on the recommended list for patient i is obtained as follows:

$$List_k^{(i)} = \begin{cases} 1 & \text{if } \varphi_k^{(i)} \in \theta^{(i)} \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

where $k = 1, 2, \dots, Q$. In other words, $List_k^{(i)}$ is 1 if patient i indeed suffers the disease recommended by the model on the next hospital visit and 0 if otherwise. Note that the half-life accuracy score for a recommendation list is $\sum_{k=1}^Q 2^{-k/c} \cdot List_k^{(i)}$, where k represents the position

on the recommended list, $2^{-k/c}$ represents the weight of the k^{th} positions on the recommended list, and c is a constant that can be adjusted as needed. Without normalization for weight $2^{-k/c}$, the half-life accuracy score is not equal to 1 even when the first G diseases on the recommended list are identical to the true values. Hence, we introduce normalization for the half-life accuracy score, i.e., the adapted normalized weight function becomes $2^{-k/c} / \sum_{g=1}^G 2^{-g/c}$. Hence, the mean accuracy score of all the samples in a test dataset is:

$$MeanAccuracy = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^Q \frac{2^{-k/c}}{\sum_{g=1}^G 2^{-g/c}} \cdot List_k^{(i)} \quad (10)$$

where $List_k^{(i)}$ represents the value of the k^{th} position on the recommended list, G is the number of diseases that patient i actually have on the subsequent hospital visit, Q is the number of diseases recommended, N is the size of the samples in a test dataset. If all the G diseases in the test dataset appear exactly in the top G positions on a recommended list, then the accuracy score is equal to 1 for this recommended list, i.e., $\sum_{k=1}^Q \frac{2^{-k/c}}{\sum_{g=1}^G 2^{-g/c}} \cdot List_k^{(i)} = 1$, $i = 1, 2, 3, \dots, N$. The ratio of recommended lists with accuracy score equal to 1 is denoted as $Ratio_{1Score}$. In addition, precision and recall can also be used to evaluate the performance of the proposed methods [49].

5. Validation of the proposed approaches using hospital datasets

This section first shows the analytical results of the proposed approaches when they are applied to two independent real-world datasets. Then we compare the performance of our proposed methods with several baseline methods. The impact of patients' demographic characteristics on the performance of the proposed method has also been investigated.

5.1. Data preprocessing and training-test schemes

5.1.1. Description of datasets and data preprocessing

Datasets for validation were collected from two general hospitals respectively in Beijing and Shenzhen, China, which are denoted as dataset A and dataset B. Dataset A includes 52,312 inpatients from January 01st 2015 to December 31st 2015, and dataset B includes 32,260 inpatients from January 01st 2016 to November 5th 2016. Information collected regarding each hospital visit includes patient identification, the sequence number of a hospital visit, age, gender, and diagnoses. Noted that dataset A does not include patients' demographic information such as age and gender while dataset B includes patients' basic demographic information.

As our goal is to predict a list of diseases that a patient might have on the next hospital visit, a patient who has only one hospital visit is removed from the datasets. Note that diagnoses in the raw datasets are clearly described using ICD-10 codes. Based on the ICD-10 codes, we mapped the diagnoses in the datasets to the DALY disease groups D_1, D_2, \dots, D_{125} derived in Section 4.1. The processed datasets ready for model validation are summarized in Table 3.

There are 84 and 82 disease groups in the processed dataset A' and dataset B', respectively. Datasets A' and B' have 78 DALY disease groups in common, while their union set covers 88 DALY groups. It is worth mentioning that many disease groups were removed from the datasets as they are acute (i.e., resulting in no second visit) or never included in

Table 3
Summary of the processed datasets for validation.

Dataset	Number of inpatients	Number of hospital visits	Number of hospital visits per patient (mean \pm SD, range)	Number of diagnoses per visit (mean \pm SD, range)	Number of DALY groups covered
Dataset A'	7989	24,466	3.06 \pm 1.67, 2–22	4.87 \pm 2.48, 1–21	84
Dataset B'	4131	13,063	3.16 \pm 1.98, 2–17	4.99 \pm 2.83, 1–11	82

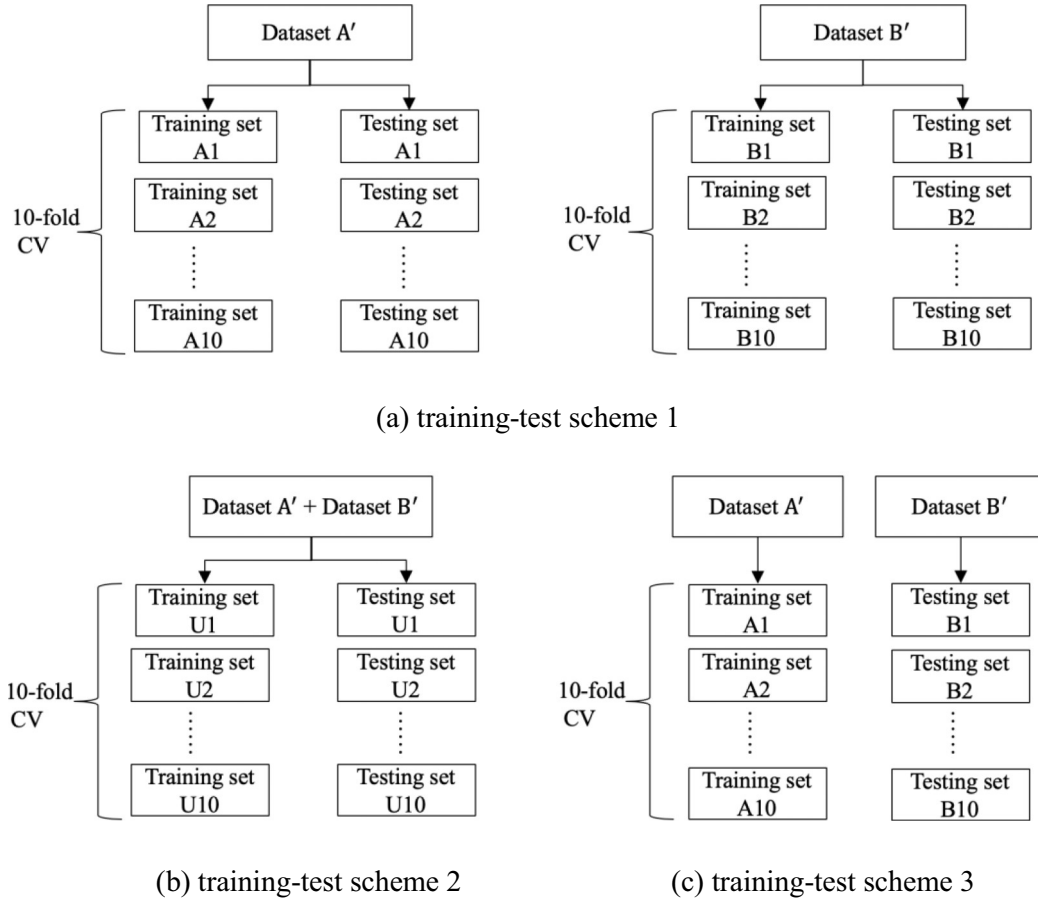


Fig. 6. Training-test schemes for datasets in the validation process.

the datasets in the first place.

5.1.2. Training-test schemes used in validation

To compare the overall accuracy of two different proposed methods, 10-fold cross-validation (CV) is applied to datasets A' and B'. As shown in Fig. 6, three different training-test schemes based on the 10-fold CV could be used in our model validation process.

The first training-test scheme (Fig. 6(a)) is to conduct the 10-fold CV on each dataset respectively. The second training-test scheme (Fig. 6(b)) is to combine dataset A' and dataset B' first and then perform the 10-fold CV on the combined dataset. The third one (Fig. 6(c)) is that dataset A' is used as the training set and dataset B' is used as the testing set, or vice versa. Based on the information about datasets in Table 3, the DALY disease groups covered by those two hospitals are different. Therefore, the third training-test scheme is not appropriate in this study.

5.2. Validation results of the proposed methods

In this study, we set $Q = 10$. Table 4 shows some examples of the lists predicted using ADTLM. In Table 4, the first column represents

patients' identification. The second column shows true value set, which is the corresponding disease groups of the true diagnoses for a patient on the next visit. For each true value set, elements are in no particular order. The third column of Table 4 provides ordered predicted lists. In other words, the predicted diseases are sorted using the descending order of risk scores. Obviously, if the position of a disease in a true value set is much closer to the top of the corresponding recommendation list, the better result it obtains in the prediction (i.e., reflecting a more accurate disease risk assessment on the list).

Using training-test Scheme 1 and 2, we have obtained accuracy scores of the recommended lists in all the training-test runs. Table 5 shows the ratio of recommended lists with 1 score and the means of accuracy scores of all the recommended lists.

As shown in Table 5, if scheme 1 is used, $Ratio_{1Score}$ gets improved by approximately 6.97% and 15.50% for dataset A' and dataset B' respectively when ADTLM is compared with DTLM. The mean accuracy score using ADTLM is higher than using DTLM by 1.56% with dataset A' and 5.13% with dataset B'. If scheme 2 is applied, $Ratio_{1Score}$ using ADTLM is also higher than using DTLM by about 10.49%, while the

Table 4
Examples of the true value sets and the lists predicted using ADTLM

Patients	True value set	Recommended list											
P1	{G ₁₀ }	G ₁₀	G ₁₄	G ₁₆	G ₉	G ₂₄	G ₂	G ₄₀	G ₁₃	G ₃₄	G ₄₈		
P2	{G ₁₆ , G ₄₉ }	G ₄₉	G ₁₆	G ₉	G ₂	G ₂₄	G ₂₂	G ₄₈	G ₄₀	G ₁₁	G ₁₃		
P3	{G ₄ , G ₁₄ , G ₃₄ , G ₄₀ , G ₆₀ }	G ₂₄	G ₄₀	G ₃₄	G ₄	G ₆₀	G ₈₄	G ₁₄	G ₉	G ₇₃	G ₄₆		
P4	{G ₉ , G ₁₆ , G ₄₉ }	G ₄₉	G ₁₆	G ₉	G ₂	G ₂₄	G ₄₈	G ₄₆	G ₁₁	G ₄₀	G ₂₂		
P5	{G ₄ , G ₁₆ , G ₃₄ }	G ₄	G ₁₆	G ₁₄	G ₉	G ₃₄	G ₂	G ₄₀	G ₁₃	G ₂₄	G ₁₁		

Table 5
Performances of the proposed approaches with two training-test schemes.

Training-test schemes	datasets	methods	$Ratio_{1Score}$ (mean \pm SD)	MeanAccuracy (mean \pm SD)
Scheme 1	Dataset A'	DTLM	0.6442 \pm 0.0086	0.8831 \pm 0.0061
		ADTLM	0.7139 \pm 0.0095	0.8987 \pm 0.0034
	Dataset B'	DTLM	0.3812 \pm 0.0176	0.8083 \pm 0.0096
		ADTLM	0.5362 \pm 0.0166	0.8596 \pm 0.0093
Scheme 2	Dataset A' + Dataset B'	DTLM	0.5511 \pm 0.0093	0.8575 \pm 0.0046
		ADTLM	0.6560 \pm 0.0088	0.8877 \pm 0.0047

mean accuracy score using ADTLM is higher than using DTLM by 3.02%. In summary, ADTLM performs better than DTLM or disease temporal link approach. The results reveal that a prediction model leveraging more patient information will perform better. Therefore, we recommend ADTLM for application in the real world.

5.3. Performance compared with other methods

5.3.1. Baselines

We compare the proposed method with three types of baseline methods, recommendation algorithms, multi-label classifiers and sequential pattern mining (SPM). Regarding recommendation algorithms, collaborative filtering (CF) is used as a baseline in this study, which had been applied to multiple disease predictive modeling [10] and is the classic benchmark method in the field of multi-disease risk prediction study [27,28]. Since the multi-disease risk prediction problem under study is different from a traditional recommendation problem, it is necessary to briefly introduce how collaborative filtering is used in the multi-disease risk prediction problem.

If a patient has disease k , then the patient has a score of 1 for disease k ; if the patient does not have disease k , then the patient has a score of 0 for disease k . The similarity between patients is calculated based on the cosine distance. Thus, the similarity between a target patient i and any other patient a is calculated as follows:

$$Sim_{i,a} = \frac{\vec{r}_i \cdot \vec{r}_a}{\|\vec{r}_i\| * \|\vec{r}_a\|} = \frac{\sum_k r_{i,k} r_{a,k}}{\sqrt{\sum_k r_{i,k}^2} \sqrt{\sum_k r_{a,k}^2}} \quad (11)$$

In Eq. (11), \vec{r}_i represents the score vector of patient i for all diseases, $r_{i,k}$ denotes the k th component of \vec{r}_i . And the dimension size of \vec{r}_i indicates the number of diseases in the recommendation system. The meaning of \vec{r}_a and $r_{a,k}$ is the same as \vec{r}_i and $r_{i,k}$.

The risk score of the target patient i suffering disease k in the future, i.e., $Score_k^{(i)}$, is calculated based on two parts [10]. One is the average prevalence of disease k , and the other is derived from medical histories of other patients who is similar to the target patient i .

$$Score_k^{(i)} = \bar{r}_k + \rho(1 - \bar{r}_k) \sum_{a \in Set_k} Sim_{i,a} \quad (12)$$

In Eq. (12), \bar{r}_k represents the average prevalence of disease k in the training dataset, Set_k represents the set of patients with disease k in the training dataset, ρ is the normalization coefficient, $\rho = 1 / \sum_{a \in Set_{train}} Sim_{i,a}$, and Set_{train} represents the set of all the patients in the training dataset. According to the recommendation list generation described in Section 4.5, the risk scores of diseases for the target patient i are sorted, then a list of ordered diseases is obtained. Finally, the first Q diseases are recommended to patient i .

From the perspective of classification, the problem in this study is a multi-label classification problem. Therefore, two classical multi-label classifiers are used to compare, i.e., K-nearest neighbor (KNN) [50] and decision tree (DT) [51]. Since the output of a classifier is a list of probabilities of developed diseases, we first sort these probabilities using a descending order and then generate an ordered recommendation list. Finally, the first Q diseases are recommended to patient i .

5.3.2. Performance compared with baselines

The performance comparisons between the proposed method and the baseline methods are shown in Tables 6–8. Note that all the results are obtained from 10-fold CV experiments.

Table 6 shows the 10-fold CV results derived from dataset A'. The proposed method ADTLM performs better than collaborative filtering by about 7.0%, 20.0%, 1.1%, 2.2% and 1.7% with respect to the metrics, *MeanAccuracy*, *Ratio_{1Score}*, *Precision*, *Recall*, and *F1 Score*, respectively. Compared to K-nearest neighbor, ADTLM has about 1.0%, 3.0%, 1.3%, 4.8%, and 2.0% improvement with respect to those metrics, respectively. ADTLM also performs better than decision tree by about

1.5%, 1.3%, 1.2%, 3.0%, and 1.8% with respect to those metrics, respectively. Compared to sequential pattern mining, ADTLM has about 4.4%, 32.5%, 4.0%, 11.0%, and 6.2% improvement with respect to those metrics, respectively.

Table 7 provides the 10-fold CV results derived from dataset B'. The proposed method ADTLM performs better than collaborative filtering by about 7.5%, 25.3%, 3.5%, 8.2%, and 5.0% with respect to the metrics *MeanAccuracy*, *Ratio_{1Score}*, *Precision*, *Recall*, and *F1 Score*, respectively. Compared to KNN, ADTLM has about 0.1%, 2.1%, 1.2%, 4.3%, and 1.9% improvement with respect to those metrics, respectively. ADTLM also performs better than decision tree by about 2.6%, 2.3%, 2.2%, 5.1%, and 3.1% with respect to those metrics, respectively. Compared to sequential pattern mining, ADTLM has about 2.6%, 31.6%, 6.0%, 14.5%, and 8.6% improvement with respect to those metrics, respectively.

Table 8 has the results derived from the combined dataset. The proposed method ADTLM performs better than collaborative filtering by about 6.7%, 21.8%, 2.2%, 5.3%, and 3.3% with respect to the metrics *MeanAccuracy*, *Ratio_{1Score}*, *Precision*, *Recall*, and *F1 Score*, respectively. Compared to K-nearest neighbor, ADTLM has about 1.0%, 3.5%, 1.4%, 4.9%, and 2.2% improvement with respect to those metrics, respectively. Again, ADTLM performs better than decision tree by about 1.7%, 1.6%, 1.4%, 3.4%, and 2.1% with respect to those metrics. Compared to sequential pattern mining, ADTLM has about 4.3%, 31.4%, 5.2%, 13.1%, and 7.9% improvement with respect to those metrics, respectively.

In summary, ADTLM performs better than the compared baseline methods, including collaborative filtering, K-nearest neighbor, decision tree, and sequential pattern mining in this study.

5.4. The impact of patients' demographic characteristics

To further explore the impact of patients' demographic characteristics on the prediction performance of the proposed methodology, patients are classified into subgroups based on age and gender. Note that dataset B' contains the demographic information of patients. Fig. 7 shows the distribution of patients by age and gender in dataset B'.

5.4.1. Grouping by gender

Table 9 gives the data summary of dataset B' grouped by gender. We perform the proposed ADTLM on these two patient subgroups using the 10-fold CV, the results are shown in Table 10.

As shown in the 2nd and 3rd column in Table 10, the performance on the female subgroup is superior to that on the male subgroup by 1.5%, 9.1%, and 1.3% in terms of *MeanAccuracy*, *Ratio_{1Score}* and *Recall*, while the male group shows a better performance on *Precision* and *F1 Score*.

To investigate if it affects the overall predictive performance on a dataset clustered by gender, the last two columns in Table 10 provide the weighted average performance of subgroups and the 10-fold CV performance without grouping. The differences are 0.0042, 0.0053, 0.0008, 0.0032, and 0.0019 for the used metrics and the differences are within the standard deviations, which shows that clustering by gender has no substantial impact on the overall predictive performance on dataset B'.

5.4.2. Grouping by age

According to the age grouping standard of WHO (0–45, 45–49, 50–59, ≥ 60) and the age structure of population (0–14, 15–64, ≥ 65) provided by China statistical yearbook, we divide dataset B' into four groups: 0–14, 15–44, 45–64, ≥ 65 by considering the age distribution in dataset B'. Table 11 gives the data summary of dataset B' grouped by age.

As there is a small number of patients in the subgroup where patients were less than 14 years old (Table 11), this age group is thus excluded from the following comparison. There are 69, 71, and 72

Table 6
Comparison with the baseline methods: Dataset A' (training-test scheme 1).

Metrics (mean \pm SD)	CF	KNN	DT	SPM	ADTLM
<i>MeanAccuracy</i>	0.8292 \pm 0.0030	0.8889 \pm 0.0052	0.8833 \pm 0.0041	0.8546 \pm 0.0048	0.8987 \pm 0.0034
<i>Ratio_{1Score}</i>	0.5138 \pm 0.0110	0.6837 \pm 0.0136	0.7013 \pm 0.0093	0.3890 \pm 0.0127	0.7139 \pm 0.0095
<i>Precision</i>	0.1943 \pm 0.0028	0.1930 \pm 0.0035	0.1942 \pm 0.0036	0.1654 \pm 0.0032	0.2057 \pm 0.0035
<i>Recall</i>	0.9238 \pm 0.0048	0.8972 \pm 0.0068	0.9153 \pm 0.0029	0.8359 \pm 0.0054	0.9456 \pm 0.0038
<i>F1 Score</i>	0.3211 \pm 0.0038	0.3177 \pm 0.0048	0.3204 \pm 0.0048	0.2762 \pm 0.0049	0.3379 \pm 0.0046

DALY disease groups covered by the last three age subgroups of “15–44”, “45–64” and “ ≥ 65 ”, respectively. It is worth mentioning that in total there are 80 disease groups covered by the last three age subgroups, while there are only 59 common disease groups across these three age subgroups. This finding clearly indicates that patients within different age groups suffer different diseases.

The proposed ADTLM was performed on the last three age subgroups using the 10-fold CV. As shown in the 2nd, 3rd, and 4th column in Table 12, the performance on the subgroup “15–44” is superior to the performance on other two subgroups in terms of *MeanAccuracy*, *Ratio_{1Score}* and *Recall*, respectively, while the subgroup “ ≥ 65 ” shows a better performance on *Precision* and *F1 Score*.

To investigate if it affects the overall predictive performance on a dataset clustered by age, the last two columns in Table 12 show the weighted average performance of age subgroups and the 10-fold CV performance without grouping. The differences are 0.0048, 0.0073, 0.0013, 0.0036, and 0.0043 for the used metrics and the differences are again within the standard deviations, which means that clustering by age has no substantial impact on the overall predictive performance on dataset B'.

In summary, there indeed exists a small different prediction performance on subgroups when clustering by gender or age. However, whether patients are grouped by gender/age or not has little impact on the overall predictive performance, which shows the robustness of the proposed method.

6. Discussions

In this section, we first discuss how the proposed framework can be appropriately employed in facilitating the decision-making process in a hospital setting. Then we summarize the contributions and highlight the limitations of our study.

6.1. Promising application in the real world

Accurate disease risk assessments can be used as a reference by medical professionals to provide patients effective healthcare guidance at the point of need. For example, the proposed ADTLM can be developed as a module embedded in a hospital information system. Once it becomes part of a decision-making support tool for medical professionals, better health intervention plans than ever before can be developed for patients to improve the medical outcomes in the long run. Fig. 8 shows such a potential application scenario.

As illustrated in detail in Fig. 8, a decision support system for

multiple disease risk assessment can be developed, which consists of software modules and a database storing data derived from ICD dictionary, DALY estimates, and EHRs data. The included software modules realize the functions required by the proposed approach. When a patient is ready for discharge, the system can recommend a list of diseases ordered by risk scores based on the patient's medical history. In addition to using personal clinical experience and medical knowledge, a medical professional can reference the recommended list to adjust and confirm his/her judgement and make a health intervention regimen for the patient. The length of a recommended list can be adjusted as needed. Medical professionals could also provide their feedback on disease risk rankings, which can be used by the system to further enhance the risk score computing module over time. Promisingly, the system can help avoid certain prejudices in decision-making. It is well recognized that medical professionals' personal clinical experience and medical knowledge can vary from one to another. As the above-mentioned system learns from massive electronic health records, therefore, it generates recommendation results in a consistent and unbiased manner for a patient.

In fact, the proposed framework can be implemented in a broader manner as patients' diagnoses today are most likely encoded as ICD codes or the like. Because the approach is robust for various data sources or code schemes, the predictive results from the applications of the approach can be well used to help health policy decision-makers understand the health risks of communities so as to help them develop optimal plans for resources and fiscal allocations.

6.2. Contributions and limitations

Although directed disease networks have been exploited in this field, there is still a gap to fill. For example, the studies by Jensen et al. [6] and Kannan et al. [7] focused on using directed disease networks either to derive disease trajectories or to determine the directionality of disease trajectories for a targeted population rather than assessing disease risks for an individual patient. The developed multiple disease prediction approach in this study filled the gap by focusing on systematically assessing future disease risks for a patient.

The most important contribution of our study lies in combining directed disease network and recommendation system techniques to facilitate multiple disease prediction modeling. By doing so, we can fully incorporate massive temporal disease relations in EHRs. As a result, we can improve the performance of developed models substantially when compared to the existing approaches in the current literature. Particularly, compared with baseline methods in our

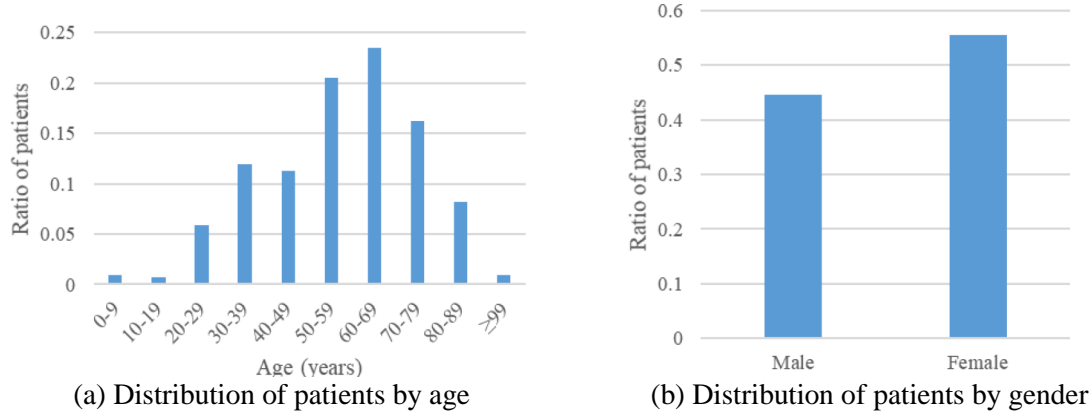
Table 7
Comparison with baseline methods: Dataset B' (training-test scheme 1).

Metrics (mean \pm SD)	CF	KNN	DT	SPM	ADTLM
<i>MeanAccuracy</i>	0.7844 \pm 0.0083	0.8584 \pm 0.0073	0.8339 \pm 0.0094	0.8340 \pm 0.0086	0.8596 \pm 0.0093
<i>Ratio_{1Score}</i>	0.2830 \pm 0.0104	0.5149 \pm 0.0163	0.5133 \pm 0.0184	0.2202 \pm 0.0110	0.5362 \pm 0.0166
<i>Precision</i>	0.2472 \pm 0.0035	0.2700 \pm 0.0037	0.2602 \pm 0.0034	0.2226 \pm 0.0043	0.2822 \pm 0.0068
<i>Recall</i>	0.8390 \pm 0.0082	0.8781 \pm 0.0051	0.8705 \pm 0.0064	0.7758 \pm 0.0090	0.9211 \pm 0.0097
<i>F1 Score</i>	0.3819 \pm 0.0042	0.4129 \pm 0.0046	0.4007 \pm 0.0043	0.3459 \pm 0.0058	0.4320 \pm 0.0077

Table 8

Comparison with baseline methods: Dataset A' + Dataset B' (training-test scheme 2).

Metrics (mean \pm SD)	CF	KNN	DT	SPM	ADTLM
<i>MeanAccuracy</i>	0.8206 \pm 0.0049	0.8779 \pm 0.0065	0.8707 \pm 0.0060	0.8447 \pm 0.0051	0.8877 \pm 0.0047
<i>Ratio_{1Score}</i>	0.4379 \pm 0.0088	0.6214 \pm 0.0153	0.6398 \pm 0.0108	0.3425 \pm 0.0094	0.6560 \pm 0.0088
<i>Precision</i>	0.2104 \pm 0.0014	0.2188 \pm 0.0039	0.2184 \pm 0.0037	0.1806 \pm 0.0012	0.2327 \pm 0.0027
<i>Recall</i>	0.8840 \pm 0.0049	0.8878 \pm 0.0058	0.9028 \pm 0.0046	0.8064 \pm 0.0050	0.9371 \pm 0.0027
<i>F1 Score</i>	0.3399 \pm 0.0018	0.3511 \pm 0.0053	0.3517 \pm 0.0049	0.2942 \pm 0.0023	0.3728 \pm 0.0034

**Fig. 7.** The distribution of patients by age and gender in dataset B'.**Table 9**

Summary of dataset B' grouped by gender

	Dataset B'	Male	Female
The number of individuals	4131	1842	2289
Age (mean \pm SD)	56.81 \pm 17.66	60.31 \pm 16.91	53.99 \pm 17.74
Average number of visits per patient (mean \pm SD)	3.16 \pm 1.98	3.11 \pm 1.90	3.19 \pm 2.04
DALY disease groups covered	82	73	77

evaluations, ADTLM has realized significantly higher *Recall* and *Ratio_{1Score}*, which are critical when predicting multiple disease risks for inpatients. Finally, the results from our conducted validations and early discussion show that the proposed methodology is robust and can be easily adopted in the real world.

Moreover, our proposed methodology improves the interpretability of the results derived from multiple disease prediction modeling. The risk score computed in our method is based on the ratios of flows among diseases, i.e., the proportion of patients diagnosed with the same disease(s), which is more explainable from the medical perspective than classifiers or similarity calculation based on the cosine distance in collaborative filtering. With two different methods proposed, i.e., DTLM and ADTLM, we prove that the more disease relationship information a model can capture, the more accurate in terms of performance it can realize.

Table 10

Prediction performance of dataset B' grouped by gender

Metrics (mean \pm SD)	Male subgroup	Female subgroup	Weighted average of subgroups	Performance without grouping
<i>MeanAccuracy</i>	0.8469 \pm 0.0113	0.8622 \pm 0.0103	0.8554	0.8596 \pm 0.0093
<i>Ratio_{1Score}</i>	0.4806 \pm 0.0147	0.5713 \pm 0.0179	0.5309	0.5362 \pm 0.0166
<i>Precision</i>	0.3063 \pm 0.0142	0.2614 \pm 0.0082	0.2814	0.2822 \pm 0.0068
<i>Recall</i>	0.9107 \pm 0.0054	0.9237 \pm 0.0066	0.9179	0.9211 \pm 0.0097
<i>F1 Score</i>	0.4582 \pm 0.0156	0.4074 \pm 0.0098	0.4301	0.4320 \pm 0.0077

One limitation of our study is that the numbers of diseases in our datasets were limited, which didn't include all the diseases described in the ICD dictionary. Another limitation is that the similarity calculation in our medical recommendation system is only based on patients' medical histories while not including their ethnicity, medication, social behaviors, etc., due to limited information in our datasets. It will be interesting to further incorporate these factors into risk score computing to understand those factors' impact on the outcomes.

While the sequence information of hospital visits is considered in our model, the time intervals between consecutive visits have not been investigated as the information is missing in our datasets for most patients. Apparently, the proposed model can be flexibly extended if richer information becomes available. It will be also interesting to explore how big data and deep learning techniques can be fully leveraged in the proposed framework to further improve the overall performance.

7. Conclusion

In this study, we proposed a novel framework that integrates directed disease networks with recommendation system techniques for multiple disease risk prediction modeling. By taking into consideration patients' disease sequence information between successive hospital visits, the proposed directed disease networks successfully incorporated temporal relations among diseases into the proposed disease risk prediction models, which makes the proposed models easy to adapt over time. We investigated two disease risk score computing approaches,

Table 11
Data summary of dataset B' grouped by age

	≤14	15–44	45–64	≥ 65
The number of individuals	48	938	1673	1472
Male:female	35:13	235:703	796:877	776:696
Age (mean ± SD)	4.54 ± 4.70	33.20 ± 6.21	55.81 ± 5.61	74.68 ± 6.99
Average number of visits per patient (mean ± SD)	2.33 ± 1.39	2.79 ± 1.64	3.48 ± 2.16	3.04 ± 1.92
DALY disease groups covered	23	69	71	72

Table 12
Prediction performance of dataset B' grouped by age.

Metrics (mean ± SD)	15–44	45–64	≥ 65	Weighted average of subgroups	Performance without grouping
<i>MeanAccuracy</i>	0.8809 ± 0.0224	0.8687 ± 0.0174	0.8225 ± 0.0157	0.8548	0.8596 ± 0.0093
<i>Ratio_{1Score}</i>	0.6933 ± 0.0490	0.5806 ± 0.0269	0.3745 ± 0.0214	0.5289	0.5362 ± 0.0166
<i>Precision</i>	0.1856 ± 0.0103	0.2648 ± 0.0042	0.3612 ± 0.0094	0.2835	0.2822 ± 0.0068
<i>Recall</i>	0.9430 ± 0.0155	0.9334 ± 0.0099	0.8901 ± 0.0064	0.9247	0.9211 ± 0.0097
<i>F1 Score</i>	0.3100 ± 0.0143	0.4125 ± 0.0054	0.5138 ± 0.0097	0.4277	0.4320 ± 0.0077

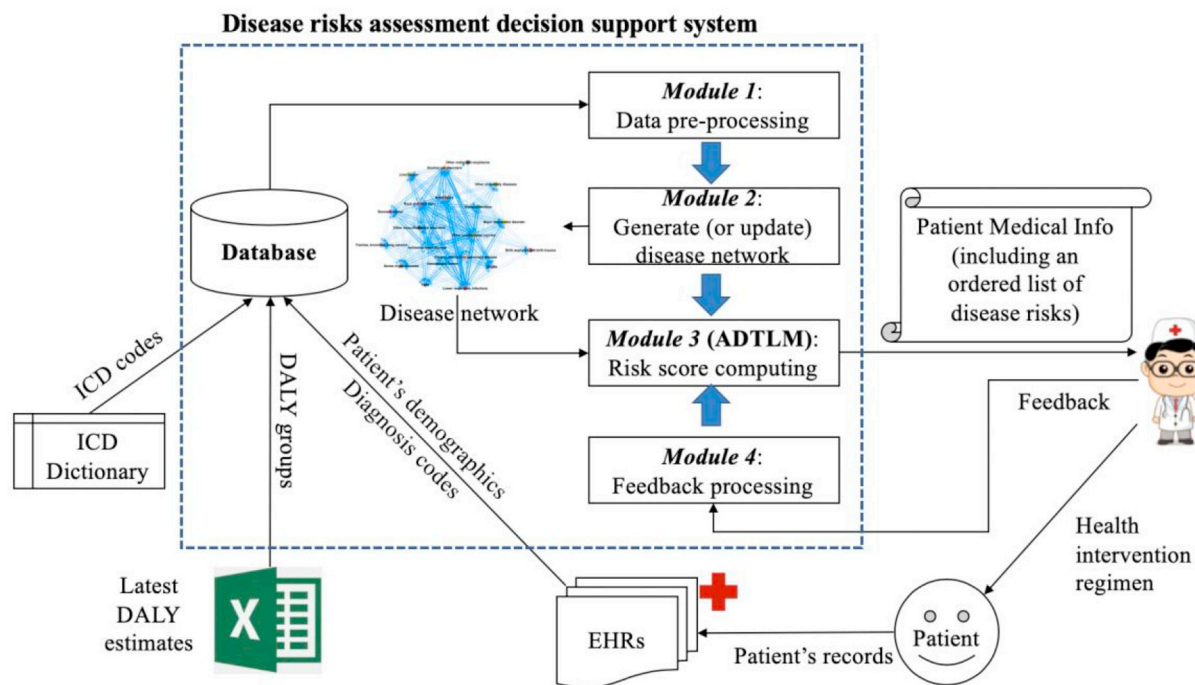


Fig. 8. Application scenario: decision support system provided to medical experts

i.e., disease temporal link approach and aggregated disease temporal links approach, to generate an ordered list of diseases risks for a patient. The proposed approaches were validated using two independent real-world datasets. The results show that the proposed aggregated disease temporal links approach or ADTLM outperforms baseline methods, such as collaborative filtering, K-nearest neighbors, decision tree, and sequential pattern mining. Additionally, we investigated the impact of patients' demographic characteristics on the performance of the proposed methodology by clustering patients into subgroups based on age or gender. The results indicate that the clustering has little impact on the overall predictive performance, which shows the robustness of the proposed methodology.

Acknowledgements

This project was partially supported by the key project of National Natural Science Foundation of China (Big Data Driven Innovation and

Management of Intelligent Healthcare, Grant No.: 71532002), IBM Faculty Awards (RDP-Qiu2016 and RDP-Qiu2017), and Penn State ICS Seed Grants (Deep Learning, 2018–19 and Reinforcement Learning, 2019–2020).

References

- [1] M. Bayati, S. Bhaskar, A. Montanari, Statistical analysis of a low cost method for multiple disease prediction, *Statistical Methods in Medical Research* 27 (8) (2018) 2312–2328.
- [2] N.V. Chawla, D.A. Davis, Bringing big data to personalized healthcare: a patient-centered framework, *Journal of General Internal Medicine* 28 (3) (2013) 660–665.
- [3] A.J. Frandsen, Machine Learning for Disease Prediction, Brigham Young University, Master thesis, 2016.
- [4] M. Nasiri, B. Minaei, A. Kiani, Dynamic recommendation: disease prediction and prevention using recommender system, *International Journal Basic Scientific Medicine* 1 (1) (2016) 13–17.
- [5] C.A. Hidalgo, N. Blumm, A.L. Barabási, et al., A dynamic network approach for the study of human phenotypes, *PLoS Computational Biology* 5 (4) (2009) e1000353.
- [6] A.B. Jensen, P.L. Moseley, T.I. Oprea, et al., Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients, *Nature Communications* 5 (2014) 1–10.

- [7] V. Kannan, F. Swartz, N.A. Kiani, et al., Conditional disease development extracted from longitudinal health care cohort data using layered network construction, *Scientific Reports* 6 (2016) 1–14.
- [8] D. Gligorijevic, J. Stojanovic, Z. Obradovic, Improving Confidence while Predicting Trends in Temporal Disease Networks, Preprint at, 2018. <https://arxiv.org/abs/1803.11462>.
- [9] K.S. Lakshmi, G. Vadivu, A novel approach for disease comorbidity prediction using weighted association rule mining, *Journal of Ambient Intelligence and Humanized Computing* (2019), <https://doi.org/10.1007/s12652-019-01217-1> Online access at.
- [10] D.A. Davis, N.V. Chawla, N. Blumm, N.A. Christakis, A.L. Barabási, Predicting individual disease risk based on medical history, *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ACM, Napa Valley, 2008, pp. 769–778.
- [11] K. Steinhauser, N.V. Chawla, A network-based approach to understanding and predicting diseases, *Social Computing and Behavioral Modeling*, Springer US, 2009, pp. 209–216.
- [12] F. Folino, C. Pizzuti, A comorbidity-based recommendation engine for disease prediction, *Proceedings of 2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, Bentley, Australia, 2010, pp. 6–12.
- [13] F. Folino, C. Pizzuti, Link prediction approaches for disease networks, *Proceedings of International Conference on Information Technology in Bio-and Medical Informatics*, Springer Berlin Heidelberg, Vienna, Austria, 2012, pp. 99–108.
- [14] A.K. Rider, N.V. Chawla, An ensemble topic model for sharing healthcare data and predicting disease risk, *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, ACM, Washington, 2013, pp. 333–337.
- [15] F. Folino, C. Pizzuti, Combining Markov models and association analysis for disease prediction, *Proceedings of International Conference on Information Technology in Bio-and Medical Informatics*, Springer Berlin Heidelberg, Toulouse, France, 2011, pp. 39–52.
- [16] F. Folino, C. Pizzuti, A recommendation engine for disease prediction, *Information Systems and E-Business Management* 13 (4) (2015) 609–628.
- [17] M. Khalilia, S. Chakraborty, M. Popescu, Predicting disease risks from highly imbalanced data using random forest, *BMC Medical Informatics and Decision Making* 11 (1) (2011) 1–13.
- [18] C.D. Chang, C.C. Wang, B. Jiang, Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors, *Expert Systems with Applications* 38 (5) (2011) 5507–5513.
- [19] A. Maxwell, R. Li, B. Yang, et al., Deep learning architectures for multi-label classification of intelligent health risk prediction, *BMC Bioinformatics* 18 (14) (2017) 523–533.
- [20] J. Billings, I. Blunt, A. Steventon, A. Steventon, T. Georgiou, G. Lewis, M. Bardsley, Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30), *BMJ Open* 2 (4) (2012) 1–10.
- [21] S.K. Thygesen, C.F. Christiansen, S. Christensen, et al., The predictive value of ICD-10 diagnostic coding used to assess Charlson comorbidity index conditions in the population-based Danish National Registry of patients, *BMC Medical Research Methodology* 11 (1) (2011) 1–6.
- [22] World Health Organization, International statistical classification of diseases and related health problems, 10th revision, Retrieved January 8, 2017. Available at, 2016. <http://apps.who.int/classifications/icd10/browse/2016/en>.
- [23] C.S. Dangare, S.S. Apte, Improved study of heart disease prediction system using data mining classification techniques, *International Journal of Computer Applications* 47 (10) (2012) 44–48.
- [24] L. Chen, X. Li, Y. Yang, H. Kurniawati, Q.Z. Sheng, H.Y. Hu, N. Huang, Personal health indexing based on medical examinations: a data mining approach, *Decision Support Systems* 81 (2016) 54–65.
- [25] M. Langarizadeh, F. Moghbeli, Applying naive bayesian networks to disease prediction: a systematic review, *Acta Informatica Medica* 24 (5) (2016) 364–369.
- [26] Y. Hao, M. Usama, J. Yang, et al., Recurrent convolutional neural network based multimodal disease risk prediction, *Future Generation Computer Systems* 92 (2019) 76–83.
- [27] D.A. Martínez-Bello, A. López-Quílez, A.T. Prieto, Relative risk estimation of dengue disease at small spatial scale, *International Journal of Health Geographics* 16 (1) (2017) 31.
- [28] T.M. Loux, C. Drake, J. Smith-Gagen, A comparison of marginal odds ratio estimators, *Statistical Methods in Medical Research* 26 (1) (2017) 155–175.
- [29] D. Dasgupta, N.V. Chawla, MedCare: Leveraging medication similarity for disease prediction, *Proceedings of 2016 IEEE International Conference on Data Science and Advanced Analytics*, IEEE, Montreal, Canada, 2016, pp. 706–715.
- [30] D.A. Davis, N.V. Chawla, N.A. Christakis, A.L. Barabási, Time to CARE: a collaborative engine for practical disease prediction, *Data Mining and Knowledge Discovery* 20 (3) (2010) 388–415.
- [31] X. Ji, S. Chun, J. Geller, A collaborative filtering approach to assess individual condition risk based on patients' social network data, *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM, California, 2014, pp. 639–640.
- [32] X. Ji, S.A. Chun, J. Geller, V. Oria, Collaborative and trajectory prediction models of medical conditions by mining patients' social data, *Proceedings of 2015 IEEE International Conference on Bioinformatics and Biomedicine*, IEEE, Washington, 2015, pp. 695–700.
- [33] X. Ji, S.A. Chun, J. Geller, Predicting comorbid conditions and trajectories using social health records, *IEEE Transactions on Nanobioscience* 15 (4) (2016) 371–379.
- [34] F. Folino, C. Pizzuti, M. Ventura, A comorbidity network approach to predict disease risk, *Proceedings of Information Technology in Bio-and Medical Informatics*, Springer Berlin Heidelberg, Bilbao, Spain, 2010, pp. 102–109.
- [35] T.H. McCormick, C. Rudin, D. Madigan, Bayesian hierarchical rule modeling for predicting medical conditions, *The Annals of Applied Statistics* 6 (2) (2012) 652–668.
- [36] X. Wang, F. Wang, J. Hu, A multi-task learning framework for joint disease risk prediction and comorbidity discovery, *Proceedings of 2014 22nd International Conference on Pattern Recognition*, IEEE, Stockholm, Sweden, 2014, pp. 220–225.
- [37] R. Li, H. Zhao, Y. Lin, et al., Multi-label classification for intelligent health risk prediction, *Proceedings of 2016 IEEE International Conference on Bioinformatics and Biomedicine*, IEEE, Shenzhen, China, 2016, pp. 986–993.
- [38] R. Bellazzi, B. Zupan, Predictive data mining in clinical medicine: current issues and guidelines, *International Journal of Medical Informatics* 77 (2) (2008) 81–97.
- [39] R. Miotto, L. Li, B.A. Kidd, et al., Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, *Scientific Reports* 6 (2016) 1–10.
- [40] E. Choi, M.T. Bahadori, A. Schuetz, et al., Doctor ai: Predicting clinical events via recurrent neural networks, *Proceedings of the 1st Machine Learning for Healthcare Conference*, 56 PMLR, 2016, pp. 301–318.
- [41] N. Razavian, J. Marcus, D. Sontag, Multi-task prediction of disease onsets from longitudinal laboratory tests, *Proceedings of the 1st Machine Learning for Healthcare Conference*, 56 PMLR, 2016, pp. 73–100.
- [42] Y.J. Kim, Y.G. Lee, J.W. Kim, et al., High Risk Prediction from Electronic Medical Records Via Deep Attention Networks, Preprint at, 2017. <https://arxiv.org/abs/1712.00010>.
- [43] F. Ma, R. Chitta, J. Zhou, et al., Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks, *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Canada, 2017, pp. 1903–1911.
- [44] P. Nguyen, T. Tran, S. Venkatesh, R. Resnet: A recurrent model for sequence of sets with applications to electronic medical records, *Proceedings of 2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Brazil, 2018, pp. 1–9.
- [45] T. Wang, R.G. Qiu, M. Yu, Multiple-disease risk predictive modeling based on directed disease networks, 2019 INFORMS Conference on Service Science, INFORMS, Nanjing, China, 2019.
- [46] A. Chen, K.H. Jacobsen, A.A. Deshmukh, et al., The evolution of the disability-adjusted life year (DALY), *Socio-Economic Planning Sciences* 49 (2015) 10–15.
- [47] World Health Organization, Estimated DALYs by cause, sex and WHO member state: 2016, Available at, 2018. https://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html.
- [48] J.L. Herlocker, J.A. Konstan, L.G. Terveen, et al., Evaluating collaborative filtering recommender systems, *ACM Transactions on Information System* 22 (1) (2004) 5–53.
- [49] A. Gunawardana, G. Shani, A survey of accuracy evaluation metrics of recommendation tasks, *Journal of Machine Learning Research* 10 (2009) 2935–2962.
- [50] M.L. Zhang, Z.H. Zhou, A review on multi-label learning algorithms, *IEEE Transactions on Knowledge and Data Engineering* 26 (8) (2014) 1819–1837.
- [51] A. Clare, R.D. King, Knowledge discovery in multi-label phenotype data, *European Conference on Principles of Data Mining and Knowledge Discovery*, Springer, Berlin, 2001, pp. 42–53.

Tingyan Wang received a Ph.D. in Management Science and Engineering from Tsinghua University in June 2018. She is now a postdoctoral researcher in the Nuffield Department of Medicine at the University of Oxford. She is currently also appointed in an honorary capacity as Data Analyst by Oxford University Hospitals NHS Trust. Most of her work has been focused on patients' electronic health records analysis and risk predictive modeling by leveraging statistics, machine learning and deep learning techniques. Her research interests include Healthcare Analytics, Medical Informatics, Disease Risk Prediction, Longitudinal Data Analysis, and Patient Care Process Analysis.

Robin Qiu holds a Ph.D. in Industrial Engineering and a Ph.D. (minor) in Computer Science both from The Pennsylvania State University. He is currently Director of Big Data Lab and Professor of Information Science at Penn State. He has had over 180 peer-reviewed publications, including 3 books. He is on the advisory board of *Service Science* and serves as an associate editor of *IEEE Transactions on Systems, Man and Cybernetics* and *IEEE Transactions on Industrial Informatics*. He was the Editor-in-Chief of *Service Science* and the Editor-in-Chief of *International Journal of Services Operations and Informatics*. He was the founding chair of the Logistics and Services Technical Committee, IEEE Intelligent Transportation Systems Society and the founding chair of Service Science Section of the INFORMS. His research interests include Big Data, Data Analytics, Smart Service Systems, Service Science, Service Operations and Management, Information Systems, and Manufacturing and Supply Chain Management.

Ming Yu holds a Ph.D. in Industrial Engineering from the National University of Ireland. He is currently Associate Professor in the Department of Industrial Engineering at Tsinghua University. He has been studying how to use industrial engineering especially information technology to the healthcare field from 2004 and he has intensive cooperation with multiple hospitals in China. He has publications in *Journal of biomedical informatics*, *BMC Medical Informatics*, and *Decision Making*. He has been working on the coding system development of the ICD-10 simplified version applied in Asia Pacific area for WHO Family of International Classifications (WHO-FIC) from 2015. He was an active member of the Program Board in several international conferences. He has been a Professional Member in Chinese Mechanical Engineering Association (CMES), and Institute Engineering, Ireland (MIEI). His research interests include Medical Informatics, Natural Language Processing, System Engineering, and Management Information System.

Runtong Zhang was born in November 1963, in Chaoyang, Liaoning, China. He got his

Ph.D. in Production Engineering and Management from Technical University of Crete in Greece in 1996, and his B.S. in Computer Science and Automation from the Dalian Maritime University in China in 1985, respectively. He is presently a professor and head of the Department of Information Management at Beijing Jiaotong University, China. He was also with the Swedish Institute of Computer Science as a senior researcher, and the

Port of Tianjin Authority as an engineer. His current research interests include big data, health-care management, operations research and artificial intelligence. He has published over 300 papers in referenced journals and conferences, and 40 books. He has been a PI for over 100 research projects and is a holder of 9 patents. He has been Senior Member, IEEE and a general chair or co-chair for over 10 IEEE sponsored international conferences.