IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 10, 2022, accepted February 22, 2022, date of publication February 25, 2022, date of current version March 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3154806

Visual Exploration of Regional Factors of the Health of Urban Residents

TANGYUAN ZOU, SONG WANG, YIFAN ZHANG, YING ZHONG, SHIJIE CHEN, AND CHAO FU

School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621000, China

Corresponding author: Song Wang (wangsong@swust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61802320 and Grant 61872304, and in part by the PhD Research Foundation of Southwest University of Science and Technology under Grant 19zx7144.

ABSTRACT Urban healthcare operations continuously generate large amounts of electronic medical record data. Combined with urban regional factors, a comprehensive data model is established to conduct a multidimensional visual analysis, which is helpful in extracting and analyzing the correlations between regional factors and residents' health. This paper establishes data models for regional factors such as regional diet, industrial economy, climate, and regional epidemics and introduces an Apriori association algorithm to find disease association relationships, a one-mode projection algorithm for bipartite networks to perform disease clustering, and a collaborative filtering recommendation algorithm to predict patients' potential high-risk diseases. Combined with the general analysis process for the visual exploration of regional health impact factors proposed in this paper, a visual analysis system is designed, and through the collaborative interaction of multiple visual interfaces, the multidimensional visual presentation of the association pattern between regional factors and residents' health is achieved. Taking a sample size of 24,626 electronic medical records in a city as an example, the case study proves that the exploration of regional factors such as catering, climate and regional epidemics can provide an analytical basis and decision support for improving the health level of urban residents.

INDEX TERMS Regional factors, urban health, association pattern, disease prediction, visual analytics.

I. INTRODUCTION

The level of social civilization and the health needs of citizens are closely related, and demand for higher-quality health services is not only a strong driving force of China's economic transformation and upgrading but also an important symbol of economic and social development [1]. Improving the health level of urban residents is in line with the requirements of China's economic transition from the high-speed growth stage to the high-quality development stage [2]. The health of the population is usually related to a variety of external factors, such as the sanitary conditions of the living environment, local climate, living and eating habits, education level and income level [3]. By exploring the correlation between external factors and residents' health, we can identify the source of health hazards and help improve the health of residents.

Electronic medical record data generated by urban healthcare facilities often implicitly characterize the health status of the local population. However, due to the characteristics of discrete data items and the subjective content and redundant structure of electronic medical record data, data processing can be difficult, making it difficult for analysts to explore the data effectively [4]. Therefore, effectively mining the hidden patterns in electronic medical record data is a current research challenge. Recent research has focused on the mining of disease association patterns [5] and clinical aid treatment [6]. For example, a study by Estruch *et al*. found an inverse association between adherence to the Mediterranean diet and the risk of cardiovascular disease [7]; Jadsri *et al*. used disease map visualization to present the spatial distribution of pollutants and patient groups in industrial areas and verified the association [8]; Kelly and Fussell. summarized the association between urban air pollution and the incidence of respiratory diseases [9]. However, these works are limited to the analysis of associations between individual regional factors and residents' health, and there is a lack of joint analysis of multiple regional factors in a macro-level and systematic way.

To address these issues, this paper proposes a common analysis process to explore the association pattern between regional factors and urban residents' health and then

The associate editor coordinating the review of this manuscript and approving it for publication was Rashid Mehmood.

designs and implements a visual analysis system based on browser/server (B/S) architecture. Based on the electronic medical records of urban medical institutions, combined with data related to food and beverages, the industrial economy, climate and other relevant areas, a comprehensive data model of urban regional factors is established to visualize the overall health situation of urban residents and to simultaneously monitor the status of multiple regional factors in multiple respects, which helps to explore the association between multiple regional factors and residents' health at the macroscopic scale. Finally, the case studies illustrate the positive impact of this work on improving the health of urban residents and promoting research in health-related fields.

The rest of this paper is organized as follows: Section 2 reviews the work related to the visualization and analytics of healthcare data. Section 3 summarizes the analytical tasks and analytical process. The methods for analyzing the association between regional factors such as regional diet, industrial economy, climate, and regional epidemics and population health are presented in Section 4. Section 5 describes the functions of the visual analytics system in detail. For each correlation analysis method proposed in Section 4, a corresponding case study is presented in Section 6 to show the effectiveness of our correlation analysis method. In addition, the main functions of the system were evaluated with a questionnaire, and expert interviews were conducted in this paper; the evaluation and interview results are analyzed and discussed in Section 7.

II. RELATED WORK

A. MEDICAL DATA ANALYTICS

The use of data mining methods on large-scale electronic medical records to find association patterns in diseases can aid clinical care, help patients prevent associated diseases and promote research in the medical field [10]. Kwon *et al*. introduced seamless integration of model parameters and hidden Markov model results into the visualization system to enable exploration of chronic disease evolutionary patterns [5]. Glueck *et al*. proposed a differential hierarchical comparison algorithm to analyze phenotypes between patients two by two, revealing the obscure hierarchical relationships in existing text-based interfaces [11]. Kwon *et al*. applied recurrent neural networks to electronic medical records and provided a visual analytic solution to the interpretability challenge of their black-box nature to achieve effective prediction of heart failure and cataract disease risk [12].

Exploring the relationship between disease and other factors such as diet and environment can help patients make positive changes to their lifestyle habits and assist health authorities in making effective decisions [13]. Estruch *et al.* designed a randomized trial to test the preventive effect of the Mediterranean dietary pattern on cardiovascular disease, demonstrating that adherence to the Mediterranean dietary pattern reduces the incidence of major cardiovascular diseases [7]. Jadsri *et al*. used disease distribution maps and pollutant dispersion heat maps to present the geographical relationship between respiratory disease patient groups and pollutants such as sulfur dioxide, and used spatial regression and time series to analyze the relationship between disease incidence and pollutants, demonstrating a significant relationship between concentrations of pollutants such as sulfur dioxide and respiratory disease incidence [8]. Kelly and Fussell. summarized the harmful effects of air pollution on the respiratory system based on epidemiological and toxicological studies, ranging from the worsening and progression of asthma, allergic rhinitis and chronic obstructive pulmonary disease to lung infections, bronchiectasis and cystic fibrosis [9]. Abdullahi *et al*. used an adaptive neuro-fuzzy inference system to predict the increase in confirmed cases of new crown pneumonia based on climate information [14].

B. MEDICAL DATA VISUALIZATION

The visualization of abstract medical text and numerical information can effectively improve the efficiency of doctors and reduce the difficulty for patients in understanding medical diagnostic data [15]. Loorak *et al*. used a combination of stacked bar charts and heterogeneous embedded data attributes to visualize time-series events and their multiple attribute types in stroke patients to improve the efficiency of the stroke treatment process [16]. Bernard *et al*. proposed a visualization technique to segment and visualize a patient's medical history as a static dashboard, enabling horizontal and vertical exploration of the patient's medical history [17]. Elshehaly *et al*. designed and implemented a dashboard generation engine by mapping the sequence of user tasks in healthcare quality improvement to a view combination unit to address the ease of use of a large-scale medical data visual analytics system [18]. Zhang *et al*. proposed a framework consisting of multiple visual representations in synergy, using a visualization approach combining body maps and sunburst diagrams to display the full range of patient health information, significantly reducing the amount of practice and effort required to access patient medical records [19].

Current research efforts are limited to mining the association between individual regional factors and health, or they lack effective integration with visual representation. To address the above shortcomings, this paper proposes a common analysis process for exploring the association patterns between regional factors and urban residents' health, and it combines various visualization and interaction methods to explore and analyze the association patterns between regional factors and diseases from multiple perspectives.

III. METHOD

A. ANALYTICAL TASKS

To explore and analyze the association patterns between regional factors and the health of urban residents, the following visual analysis tasks are proposed after close discussions with experts in the field regarding the current problems

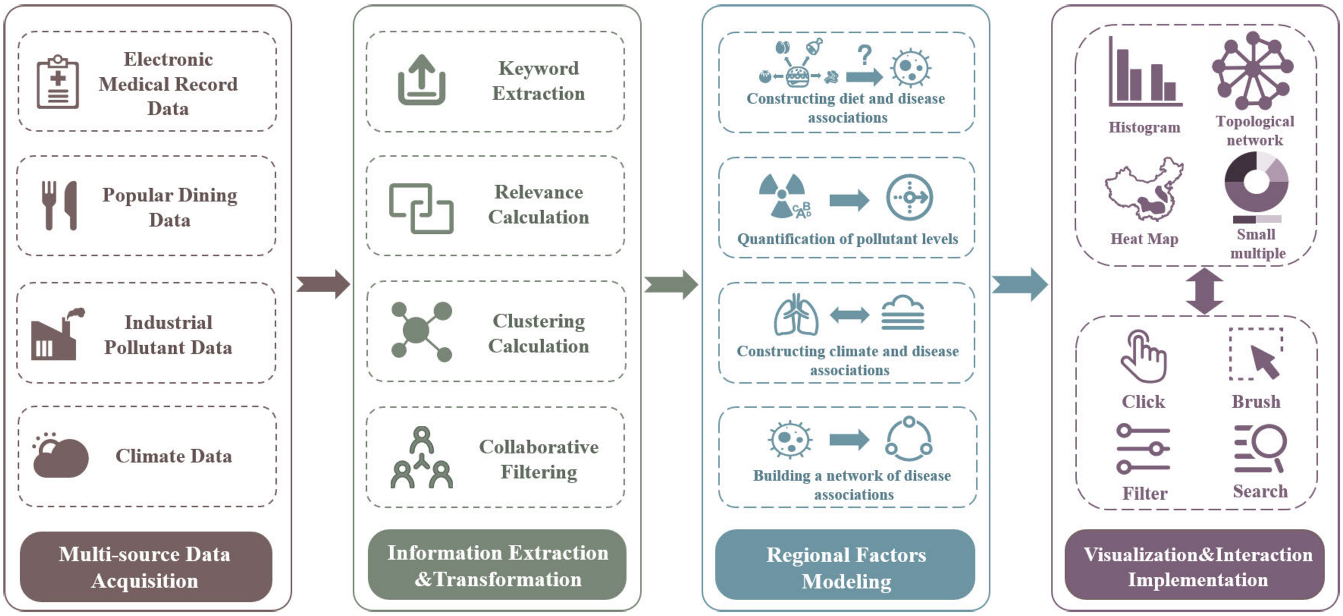


FIGURE 1. **A generic analysis process for exploring regional factor associations. Potential association patterns between regional factors and residents' health are mined through data acquisition, data processing and transformation, model building for each regional factor, and visualization.**

and shortcomings in the management of the health of urban residents:

- **T1** **A large screen presents an overview of the health status of the city's inhabitants and the spatial and temporal distribution of medical resources.** This will be used to assist experts in the field in analyzing the health characteristics of a city's inhabitants, adjusting the allocation of health resources and rationalizing the planning of public facilities.
- **T2** **Exploring the association between health and regional factors such as dietary habits, industrial economy and seasonal climate.** Understanding the patterns of association between residents' health and external regional factors can help in developing a multifaceted approach to improving the health of the population.
- **T3** **Disease association mining and analysis.** Exploring disease association patterns can help prevent complications in advance, shorten the course of patients' illnesses and improve clinical efficiency.
- **T4** **Disease popularization and prediction of potential high-risk diseases.** This is a way to raise the level of medical awareness of the population and help people understand their own health status through disease awareness. At the same time, it will help in predicting the potential risk of diseases based on a patient's disease status, personal habits and living environment.

B. ANALYTICAL PROCESS

To achieve the task of exploring the association between regional factors such as diet, industrial economy, climate and regional epidemics and the health of the population, this paper proposes a common analytical process for exploring and analyzing the association between regional factors and the health of urban residents. As shown in Figure 1, the process consists of 4 steps.

(1) Multisource data acquisition. The model associating between regional factors and population health includes data on regional factors such as food and beverage, industry, economy, and climate, in addition to the key urban electronic medical record data. For restaurant data, the names of restaurants ranked by hotness and the corresponding restaurant types and user ratings are crawled from dianping.com; for industry and economy-related data, city GDP and key polluting enterprises are obtained from City Open Data; and for climate data, city temperature, air quality, rainfall and other data are obtained from China Weather.

(2) Information extraction and transformation. The TF-IDF algorithm is used to extract key information from the raw data, the Apriori association algorithm is used to calculate disease associations, the one-mode projection algorithm for bipartite networks is used to calculate disease clustering, and the collaborative filtering algorithm is used to calculate potential high-risk diseases for patients. All the processed data are aggregated according to each regional factor and the corresponding data warehouse is built separately.

(3) Modeling of regional factors. Various models, such as models of the association between eating habits and health, pollutants and health, climate and health, and diseases are constructed.

(4) Interactive visual presentation. Visualizations such as histograms, topology diagrams, heat maps and small

multiples are used to present patterns of association between regional factors and the health of the population, combined with interactive methods such as boxing, filtering and searching.

IV. VISUAL ANALYSIS OF REGIONAL FACTORS

A. EXPLORING THE IMPACT OF REGIONAL DINING HABITS ON HEALTH

Modern medicine has demonstrated a strong correlation between dietary habits and certain chronic diseases such as cardiovascular disease, diabetes, gout, and hypertension [20]. Quantitative analysis of the relationship between diet and disease is difficult due to the complex structure of the food composition network [21], and there is no specific numerical formula in the medical literature for the association between food composition and disease [22]. To address the above problems, this paper proposes a model for the association between different dietary categories and the risk of chronic diseases to analyze the association between dietary habits and the high morbidity of chronic diseases.

The model is constructed in 4 steps:

(1) Extracting the characteristics of the dietary habits of the residents. The first step is to analyze the information on popular restaurants in the city; based on the textual sentiment analysis of user ratings and comments, we extracted the food preferences of user groups as the characteristics of the eating habits of people in the area.
(2) Deconstructing the nutritional composition of the diet. The composition of the diet of the residents is deconstructed according to the China food composition [23], and a table of diet and nutrition compositions is established.
(3) Development of a disease risk index evaluation model. Experts in the field are invited to identify evaluation indicators and judgment matrices, and to develop a model for assessing the risk due to common dietary nutrients and high-incidence chronic diseases using the analytic hierarchy process (AHP).
(4) Optimisation of the model parameters. To increase the accuracy of the evaluation model, the model is validated on medically proven cases of high-fat and high-salt diets versus Mediterranean diets [7].

B. EXPLORING THE IMPACT OF REGIONAL POLLUTION FACTORS ON HEALTH

Before analyzing the association between industrial pollution and the health of urban residents, it is necessary to model the association between extent and level of pollutants and the health of surrounding residents. Referring to the extrapolation method of the radiation radius of point source pollution proposed by Xiaolong and Sun [24], this paper establishes a radiation radius model of pollutants according to the class and type of pollutants to assess the risk level of pollutants to human health. The main construction process of the model is as follows:

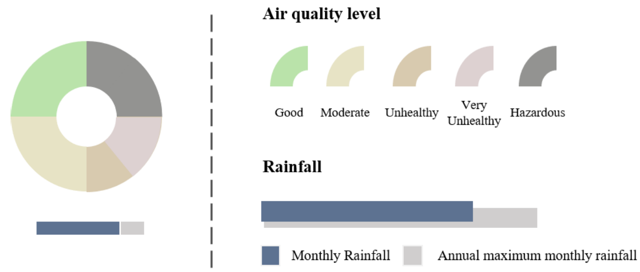


FIGURE 2. Visualization design of multiattribute aggregated climate small multiple information.

(1) Pollutant classification. According to the detection index of industrial pollutants, pollutants are initially classified into three categories: air pollution, solid waste pollution and water pollution.
(2) Determining pollution levels. The pollutants are further classified according to heavy metals, chemicals, organics, etc., and the pollution risk thresholds are determined according to the pollution scale of the corresponding categories.
(3) Calculation of the radiation radius of the pollution risk. Based on the obtained pollutant emissions and the pollution risk threshold, the radiation radius of the pollution risks of enterprises is calculated by referring to the extrapolation method of the radiation radius of point source pollution.

C. EXPLORING THE IMPACT OF REGIONAL CLIMATIC CHARACTERISTIC FACTORS ON HEALTH

Climate data usually include various attributes such as air quality, rainfall and temperature, and have a temporal dimension [25]. The traditional combination of bar charts and line charts cannot effectively present the multidimensional attribute information and the correlation relationship among attributes. Therefore, in this paper, a novel multiattribute aggregated climate information small multiple is designed. In Figure 2, the air quality and rainfall information is presented in small multiples by month. The small multiples contain two parts: the scale bar graph presents the monthly rainfall situation, the dark part indicates the rainfall in the current month, and the light background part indicates the maximum rainfall of the year, which is used to unify the relative proportions of rainfall in each month and facilitate comparison and analysis. The air quality information is presented as a circular pie chart with month-specific text headings inside; the sectors of the pie chart are color-coded with air quality classes, and the number of days with a corresponding air quality class is coded with the size of the radian, thus achieving the joint presentation of overall and local climate information and the exploration of the correlation between rainfall and air quality distribution.

By arranging the small multiples of 12 months into one view, we can visually present the air quality changes for a whole year and explore the correlation between air quality trends and respiratory disease incidence jointly with the distribution of respiratory disease patients on the climate analysis page of this system (Figure 4(C3)).

D. DISEASE ASSOCIATION ANALYSIS

For the disease association mining analysis task, this paper introduces the Apriori algorithm to mine the frequent item sets and association rules of diseases, improves the bipartite graph one-mode projection algorithm and applies it to the disease data to extract the disease clustering structure, and uses a network topology diagram to visually present the association rules and clustering structure of diseases.

1) DISEASE ASSOCIATION CALCULATION

Inspired by the related work of Khedr *et al.* [26], this paper introduces the Apriori algorithm to mine the association relationship between diseases.

$$support\ (A \Rightarrow B) = P\ (A \cap B) \quad (1)$$

$$confidence\ (A \Rightarrow B) = P\ (B \mid A) = \frac{support\_count\ (A \cap B)}{support\_count\ (A)} \quad (2)$$

where $support\ (A \Rightarrow B)$ is defined as the probability of the simultaneous occurrence of A and B, $confidence\ (A \Rightarrow B)$ is defined as the probability of B occurring under the condition that A occurs, $support\_count$ is the frequency of the item set, and the core of the algorithm is to find the frequent item set of the disease combination. By observing the output of the algorithm and adjusting the values of the support and confidence lower limit parameters, the statistically significant disease association rules are selected.

2) DISEASE CLUSTERING CALCULATION

To generate the clustering structure of diseases, the disease name and the attributes of the disease (e.g., department and symptoms) are partitioned into two disjoint sets of nodes and formed into a bipartite graph. Based on the idea of the bipartite graph weighted one-mode projection algorithm of Zhou *et al.* [27], if two nodes have connected nodes with common attributes, an edge is generated between the two nodes, and the weighted sum of the similarities of all common attributes of the nodes is the distance between the two nodes. However, some attributes of different diseases (e.g., clinical manifestations and etiology) have high similarity but low overlap in text content; i.e., the common attribute nodes intersect but do not completely overlap. Therefore, the clustering structure of the disease cannot be generated directly using the one-mode projection algorithm.

In this paper, we quantify the similarity of intersecting attribute nodes by using the cosine formula and merge the intersecting attribute nodes whose similarity reaches more than 50%. To distinguish these from fully overlapping attribute nodes, the connected edges generated by the merged attribute nodes are given similarity weights. Given that different disease attributes have different degrees of portrayal of

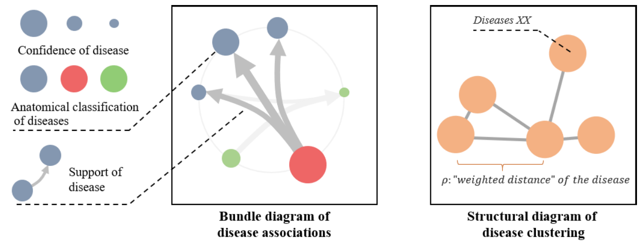


FIGURE 3. Visualization design of a disease association network.

disease characteristics, this paper adds portrayal weights to each disease attribute to improve the accuracy of the clustering results. Taking the disease nodes s and t as an example, the distance value is calculated by:

$$D\,(s,t) = \sum_{i=1}^{m} \frac{\sum_{j=1}^{n} s_{ij} \times t_{ij}}{\sqrt{\sum_{j=1}^{n} \left(s_{ij}\right)^2 \times \sum_{j=1}^{n} \left(t_{ij}\right)^2}} \times v_i \quad (3)$$

where, s_{ij} and t_{ij} denote the jth keyword frequency vector value of the text information of the ith attribute of disease nodes s and t after the word separation process, respectively, and v_i denotes the influence weight of the ith attribute of disease node s.

The core steps of the disease clustering structure generation algorithm are as follows:

(1) Merging of intersecting nodes. The similarity of each attribute of two disease nodes is calculated for intersection node merging.
(2) One-mode projection. All the connected edges generated by two disease nodes are weighted and combined according to the similarity weights and the attribute node portrayal weights to generate a set containing multiple disease clustering unipartite graphs.
(3) Disease clustering structure extraction. The graph traversal algorithm is used to traverse the set of unipartite graphs from the target disease, record the path nodes and edges, and generate the target disease clustering structure graph.

3) DISEASE ASSOCIATION VISUALIZATION DESIGN

Using the network layout algorithm to abstract the association of diseases as a node-link graph, we can effectively explore the association patterns of diseases from the perspective of topology and explore the disease features by incorporating knowledge of graph theory [28].

As shown in Figure 3, to explore the association and clustering of diseases visually, we abstract the diseases as nodes; the node size encodes the confidence of diseases, the node color encodes the anatomical classification of diseases, the support relation between diseases is abstracted as directed edges, and the association results of diseases are visualized and constructed as bundle graphs. In addition, the disease distances obtained from the clustering calculation are abstracted as edges, where the length of the edges encodes the magnitude of the weighted distance values of the disease attributes, and the set of diseases and their attributes are abstracted as nodes to visualize the disease clustering using a force-oriented graph.

When analyzing disease associations in the disease association bundle diagram, the target node, its associated nodes and its associated edges are highlighted to facilitate the rapid understanding of disease association patterns. The directions of the association edges locate the relative position of the associated disease in the course of the target disease, the width of the association edges reflects the probability of occurrence of the associated disease, and the exit and entry degrees of the target nodes reflect the prevalence pattern of the target disease with several comorbidities. In the disease clustering structure diagram, the higher the similarity of diseases is, the shorter the distance between the disease nodes, so similar diseases form class clusters, and it is possible to quickly grasp similar diseases by observing the class clusters in which the target disease nodes are located. At the same time, intermediate disease nodes connecting different classes of clusters may be crucial to preventing the spread of complications, while isolated nodes are likely to be rare diseases.

E. POTENTIAL DISEASE RISK PREDICTION

Patient groups with similar disease conditions usually share similar work environments or lifestyle habits, and the progression of the disease is usually very similar [29]. Therefore, in this paper, we introduce a user-based collaborative filtering algorithm for potential disease risk prediction for patients. The core steps of the algorithm are as follows:

(1) The degree of similarity between patients is calculated using cosine similarity based on the list of patients' diseases.
(2) A number of patients most similar to the target patients are selected, and the diseases of all patients are extracted.
(3) The target patient's diseases are removed and the remaining diseases are considered as candidate potential risk diseases for the target patient.
(4) The risk indices for all candidate diseases are calculated based on the cosine similarity of patients and the population morbidity of the disease.

The risk index of the disease is calculated by the formula:

$$P(u, i) = \sum_{v \in S(u,K) \cap N(i)} w_{uv} \times r_{vi} \tag{4}$$

where, $S(u, K)$ is the set of K patients most similar to patient u, $N(i)$ denotes the disease i suffered by patient v, w_{uv} is the cosine similarity of patient u to patient v, and r_{vi} denotes the population morbidity of disease i suffered by patient v.

The algorithm for the calculation process of potential high-risk diseases and the risk index for the specified patient u is as follows:

The collaborative filtering algorithm used in this paper is mainly based on the historical medical record data of the patient population for screening and matching risk diseases and recommendations. Individual patient histories are abstracted into disease vectors, and the degree of similarity between patients is assessed using cosine similarity. Patients with a higher degree of similarity are more likely to have similar courses of disease. Prediction of potentially risky diseases is achieved by referring to similar courses of disease.

Algorithm 1 Risk Index Calculation for Potentially High-Risk Diseases

Input: Patient u's disease list $N(u)$, set of all patient disease lists I, Patient similarity matrix T, population morbidity set for all diseases R, list of all diseases $N(u, K)$ of the K patients most similar to patient u, list of candidate potential risk diseases for patient u $M(u)$.
Output: List of potential high-risk diseases and disease risk indices S for patient u.

for *patient* i, j *in* I **do**
  $T \leftarrow cos(i, j)$ calculate disease frequency list R
end
$N(u, K) \leftarrow N(max(T[u], K))$ $M(u) \leftarrow N(u, K) - N(u)$ **for** *disease* i *in* $M(u)$ **do**
  $S \leftarrow P(u, i)$
end

The relationship between human health and external factors is complex and difficult to quantify precisely through mathematical formulas [13]. Nevertheless, the use of statistical methods to uncover the characteristics of the diseased population and the development patterns of some diseases is an important reference for medical aid diagnosis and disease prevention in patients.

V. SYSTEM OVERVIEW

In integrating the above analysis methods, this paper uses open-source JavaScript libraries such as Mapbox GL JS, d3.js and echarts.js to obtain the front-end visualization charting components and Python libraries such as numpy and pandas to complete the back-end calculation; we design and develop an interactive analysis system with multiple visualization interfaces based on B/S architecture. The system consists of three functional pages.

(1) Disease analysis page. The disease organ map (Figure 4(A1) and (A2)), the gender-age ratio of patients with high-incidence diseases (Figure 4(A7)), the disease encyclopedia (Figure 4(A8)) and the monthly overview of high-incidence diseases (Figure 4(A9)) are combined to perform disease feature mining and calculate the disease knowledge popularization functions. Disease association mapping (Figure 4(A4)), disease clustering structure mapping (Figure 4(A5)) and disease knowledge mapping (the visual exploration of common features of diseases) (Figure 4(A6)) are the views that jointly complete the disease association mining function. Figure 4(A3) shows the calculated results of potential high-risk diseases of patients predicted by the system.

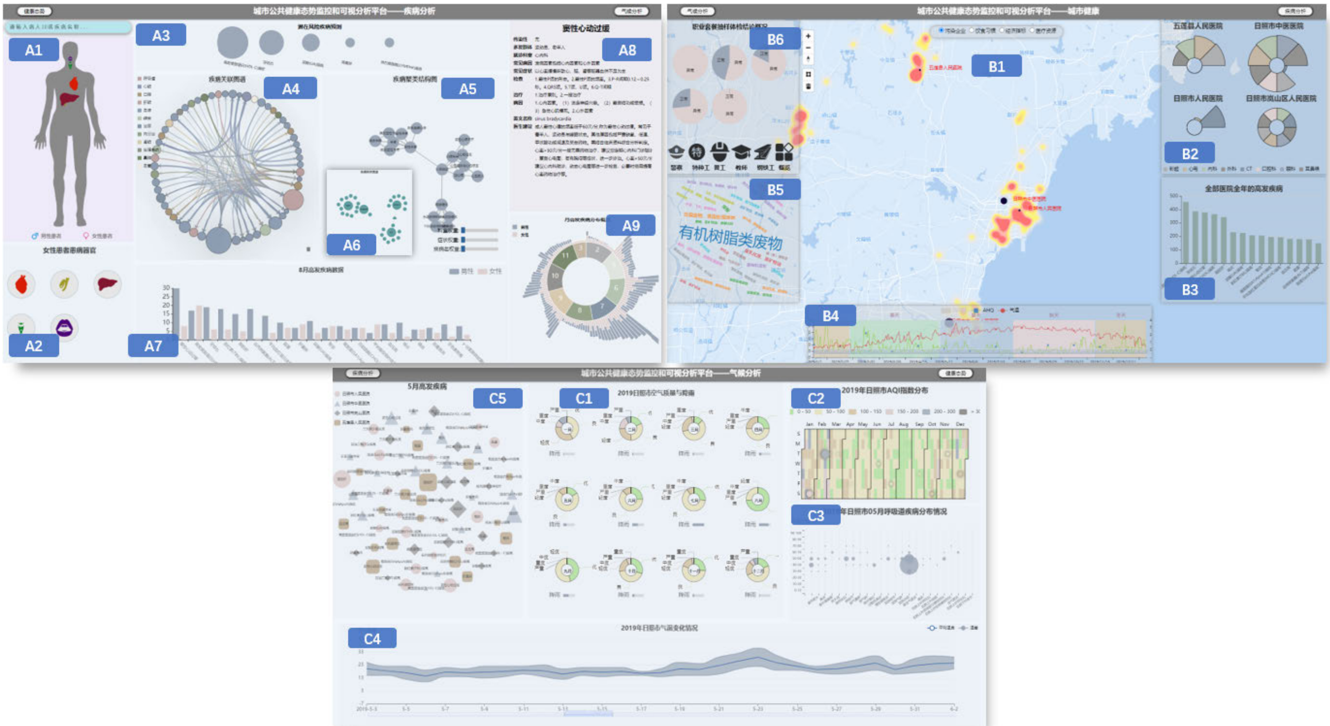


FIGURE 4. System overview. The system consists of a disease analysis page(A1-A9), a city health page(B1-B6) and a climate analysis page(C1-C5).

(2) City health page. The city map (Figure 4(B1)) shows regional factors such as the distribution of polluting enterprises and their pollution status, the distribution of popular restaurants, an overview of the distribution of enterprises contributing to the city's GDP and the percentage of their contribution, and the distribution of public health units in the city. Figure 4(B2) shows the pressure situation of visits to each department in all secondary class A hospitals in the city. Figure 4(B3) shows the high prevalence of diseases in the selected hospitals. Figure 4(B4) is used to filter the time range to be analyzed. Figure 4(B5) shows the key information within the selected analysis area in the map. Figure 4(B6) shows the health status of several typical occupations.

(3) Climate analysis page. This consists of multiattribute aggregated climate information small multiples (Figure 4(C1)), the annual air quality index (AQI) distribution (Figure 4(C2)), the distribution of respiratory disease patients (Figure 4(C3)), temperature variation (Figure 4(C4)) and the distribution of high-incidence diseases (Figure 4(C5)), which together complete the exploration of regional climate characteristics and the association with urban residents' health.

VI. CASE STUDY

This paper uses electronic medical record data of a prefecture-level city in China as a sample. The data include 6879 samples of medical examination report data and 17747 samples of examination item data from public hospitals in this city from February 2019 to November 2019, and each data sample contains data items such as the patient ID, age, gender, examination department, examination result, test time, examination hospital, and doctor dietary recommendation. In addition, we collected data on air quality, temperature and rainfall in the city in 2019, the names of 153 key polluters and the main pollutants emitted, and the names of 500 popular restaurants and information on the types of restaurants.

A. EXPLORING THE REGIONAL FACTORS OF REGIONAL DIETS

The areas near the traditional Chinese medicine (TCM) hospitals in the city with a dense restaurant distribution are selected for analysis and explored, and the gray circular range (Figure 5(A)) corresponds to the approximate address range of all patients visiting TCM hospitals in the city in 2019. From the word cloud of food and beverage categories shown in Figure 5(B), the five most popular food and beverage categories in this range are fast food, seafood, tea and juice, bread and bakery, and grilled meat. Figure 5(C) shows the distribution of disease risk indices for all the dietary categories in the range according to the association model between dietary categories and chronic disease risk in this paper, as well as a joint view of the number of patients with the most prevalent diseases in the region, which shows that the risk indices of the popular dietary categories in the region are approximately the same as the distribution of the most prevalent diseases.

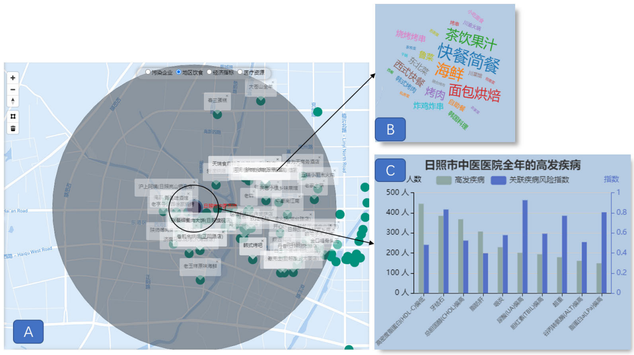


FIGURE 5. Analysis of dietary disease association near TCM hospitals.

Further analysis revealed that the number of patients with high dietary-associated disease risk indices, such as high uric acid (UA), overweight and high lipoprotein a (LPa), is relatively low, contradicting the popularity of seafood and high-fat, high-salt dishes such as barbecue and hot pot, and the survey found that the city is a tourist-oriented city with a small resident population, and seafood is the specialty of the city. It is assumed that the long-term consumers of seafood and grilled meat in the city are mainly foreign tourists, and local residents are aware of the hazards of long-term consumption of such food. The number of patients with low high-density lipoprotein (HDL-C) and fatty liver is higher than the number of those with a relatively low risk index of diet-related diseases, so it is assumed that the residents of the city are less aware of the hazards of high cholesterol and high sugar dietary habits, such as drinking alcohol, tea and juice, and eating cakes.

B. EXPLORING THE REGIONAL FACTORS OF THE INDUSTRIAL ECONOMY

The area near the People's Hospital, which has a dense distribution of high-pollution enterprises in the city, is selected for analysis and exploration, as shown in Figure 6(A1). The main types of polluters in this area are sewage treatment plants, manufacturing, food production companies and power companies. According to the discharge of pollutants from all selected enterprises shown in Figure 6(A2), the main pollutants in the area are organic resin waste, surface treatment waste, waste mineral oil and nickel-containing waste. According to the distribution of high-prevalence diseases in the region presented in Figure 6(A3), liver-related diseases such as high hepatitis B surface antibody, fatty liver, high total cholesterol and liver cysts are frequent in the region. Selecting the Wulian County area of the city for comparative analysis, we found that the main pollutants emitted by the enterprises and the high incidence of diseases in the area are approximately the same as those in the area near the People's Hospital, as shown in Figure 6(B1), (B2) and (B3).

Organic pollutants such as organic resinous waste can spread along the food chain and accumulate in body fat, eventually causing damage to the human liver [30], which may be the cause of the frequent occurrence of liver-related diseases in the region. It is inferred that the high incidence of diseases in the region is closely related to industrial pollutants.

C. EXPLORING THE REGIONAL FACTORS OF CLIMATIC CHARACTERISTICS

May, June and August are selected from among the systematic small multiples (Figure 4(C1)) to explore the association between climate features and diseases. The results in Figure 7(A1) show that the number of days in May with air quality ratings below excellent is 26, the amount of monthly rainfall is 62 mm, and the total number of patients with respiratory diseases is high, accounting for a significant proportion of all the high-incidence diseases in the month (Figure 7(A2)), including 39 cases of pharyngitis (Figure 7(A3)). Compared to May, June has a higher rainfall of 106 mm and slightly better air quality, and the number of patients with respiratory diseases is about the same, but the total number of patients with a high incidence of diseases is higher. August has a high rainfall of 178 mm, the best overall air quality of the year, with 22 days of excellent air quality, and a low number of respiratory illnesses, including 5 cases of pharyngitis. The temperature trends for the three selected months are relatively smooth, as shown in Figure 7(A4), (B4) and (C4). Figure 4(C2) shows the air pollution status of the city on each day in the form of a calendar chart. The numbers of days with excellent and good air quality in May, June and August are generally concentrated, and the air quality changes are relatively flat.

Based on the above information, it is assumed that the air quality condition of the city is affected by the amount of rainfall, the air pollution level is approximately inversely related to the amount of rainfall, and the number of respiratory patients is approximately positively related to the air pollution level.

D. ASSOCIATED DISEASE EXPLORATION AND DISEASE POPULARIZATION

Taking the disease "turbinate hypertrophy" as an example, systematic disease association mapping shows that the associated diseases include pharyngitis, fatty liver, high cholesterol and overweight, as shown in Figure 8(A). Two disease nodes, fatty liver and high cholesterol, are selected, and the disease knowledge map (Figure 8(B)) shows that fatty liver and high cholesterol share common subattributes of infectiousness, department of consultation, and multiprevalence group. The disease clustering structure map (Figure 8(C)) shows that the most similar diseases directly related to this disease are nasal polyps, rhinitis and sinusitis, which are very similar in clinical presentation and may lead to misdiagnosis. According to the Disease Encyclopedia(Figure 8(D)), one of the important causes of the disease is "chronic stimulation by alcohol and tobacco", and according to the gender and age distribution of patients with the disease shown in Figure 8(E), the main incidence group is adult males between 20 and 60 years old. A search of the system for pharyngitis and fatty liver revealed

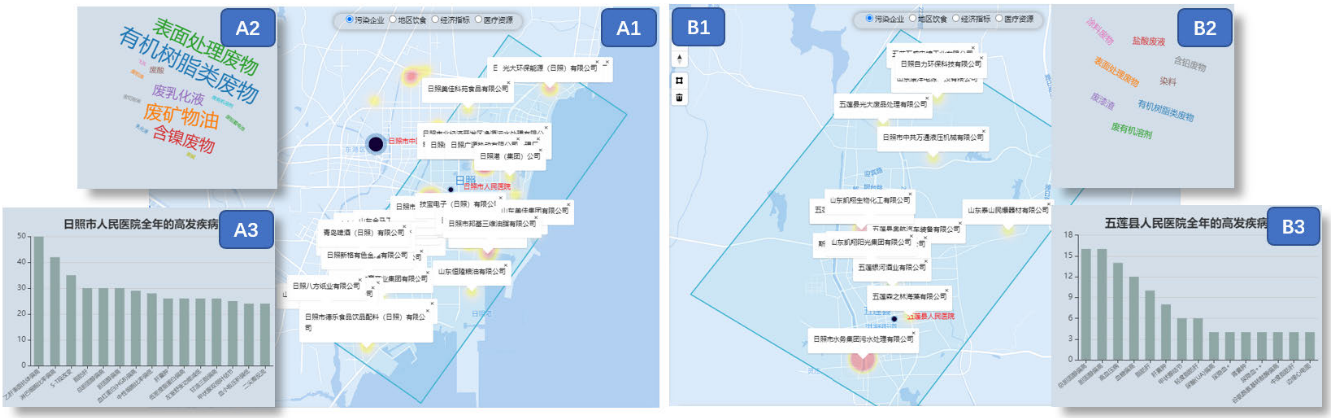


FIGURE 6. Pollutant association analysis near people's hospital and wulian county.

FIGURE 7. Exploring climate and disease associations from june to august.

that "chronic alcohol and tobacco stimulation" is also an important cause of both diseases.

Alcohol and tobacco habits are common in adult men, and these habits increase the risk of turbinate hypertrophy, pharyngitis, and fatty liver disease. In addition, fatty liver patients are often overweight and have high cholesterol. Therefore, patients with this disease should be careful to avoid smoking and alcohol and to control their weight by diet to stop the progression of the disease in advance.

E. POTENTIAL HIGH-RISK DISEASE PREDICTION

Taking the patient with medical record number "995c205a63 a19d9aab4f44ea2607b573" as an example, Figure 9(B) shows the patient's diseases, including sinus bradycardia, gallbladder polyp, fatty liver, liver cyst, pharyngitis, and dental pigmentation. The potential high-risk diseases of this patient (Figure 9(E)), in descending order of risk index,

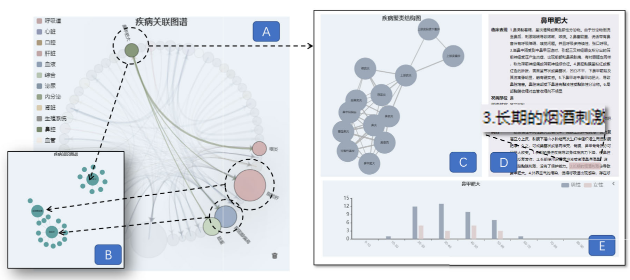


FIGURE 8. Exploring the association of "turbinate hypertrophy" with other diseases.

are low HDL-C, dental calculus, high UA, and renal cysts. Based on Figure 9(C2) and (C3) and the associated diseases shown in the disease association profile (Figure 9(D)), it is assumed that the patient is likely to suffer from obesity

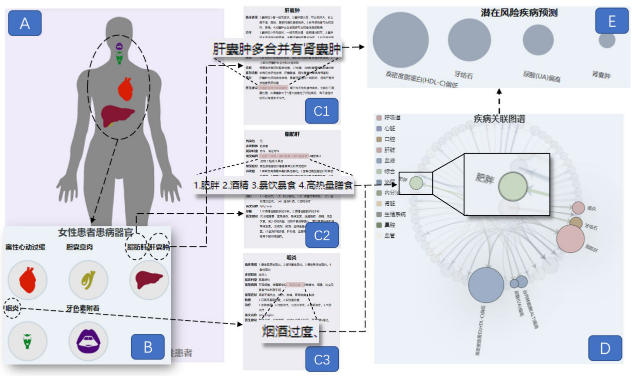


FIGURE 9. Potential high-risk disease prediction and interpretation of the prediction results.

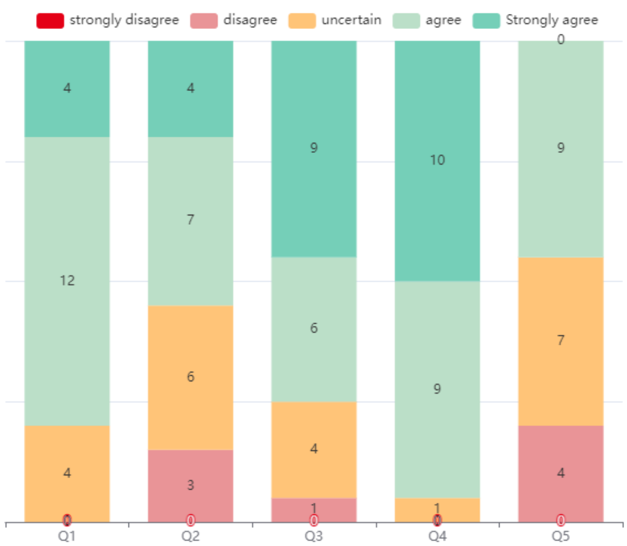


FIGURE 10. Statistical diagram of the assessment questionnaire results.

and may also have smoking, alcohol and binge eating habits.

Among the potential high-risk diseases predicted by the system, low HDL-C is common in patients with obesity, and dental calculus is more prevalent in those with smoking and alcohol habits; seafood cuisine is a specialty of the patient's region, and high UA due to the consumption of seafood is also a common disease in the region. If the patient does not pay attention to his diet, he is more likely to develop high UA. The physician's recommendation in Figure 9(C1) suggests that hepatic cysts are associated with congenital genetic factors and are often combined with renal cysts, and this patient may require ultrasound or CT to determine the health of the kidneys.

VII. EVALUATION

To further demonstrate the effectiveness of the system in uncovering patterns of association between regional factors and residents health, an evaluation questionnaire was designed, and users were invited to evaluate the system. A total of 20 volunteers were invited to participate in the evaluation, including 5 computer-related professionals, 5 medical professionals, 5 public health administrators and 5 nonspecialists.

Before evaluating the system, volunteers first received an explanation of the basic functions of the system and how to use it, and then they explored the system according to the four analysis tasks proposed in this paper and rated the questions in Table 1 based on their subjective feelings during use. This paper uses a five-point Likert scale, ranging from 1 to 5, indicating "strongly disagree", "disagree", "uncertain", "agree", and "strongly agree".

The statistical results of the evaluation are shown in Figure 10. From the analysis and evaluation results, it is concluded that the visual analysis system of this paper is excellent in presenting the overall health status of urban residents (Q1) and popularization of medical-related knowledge (Q4), and it can present the overall health status of urban residents. However, some volunteers commented that the overall interface presented too much content at the same time, which caused cognitive overload. The overall performance on Q3 was good; it can visualize the association patterns of diseases in a more comprehensive and three-dimensional manner.

The performance of the system is not particularly satisfactory in terms of exploring and analyzing the potential relationship between regional factors and population health (Q2) and predicting potential high-risk diseases (Q5). The analysis of the feedback results from the user follow-up survey shows that the system has many functions and the connection between functions is not very strong, so it is necessary to analyze the connection with the user's own expertise, and it is difficult to complete the task of potential relationship mining and analysis for users who do not have relevant expertise. In addition, half of the volunteers questioned the reasonableness of the disease prediction results given by the system. Although the principles and algorithmic processes of disease prediction are described in detail in this paper, only the prediction results are visually presented in the system interface, and there is no representation of the prediction principles or processes, resulting in a cognitive disconnect for the users, which may be the main reason why volunteers expressed doubts. Some volunteers also reported that the predictions given by the system deviated significantly from the norm when searching for patients with rare diseases. This is because the prediction results are produced based on statistical algorithms that match recommendations to existing patient records, and when the diseases of the analyzed subjects are rare, the sample size of similar patient groups matched by the algorithm is very small, leading to a decrease in the reasonableness of the recommendation results.

After completing the questionnaire, we conducted informal interviews with several experts and collected their opinions and feedback on the visual design, user interaction and usability of the system.

TABLE 1. Evaluation questionnaire for the main functions of the system.

Serial Number	*Question Description*
Q1	Use of the system can yield a clear picture of the overall health status of the city's residents
Q2	Use of the system can uncover the potential relationship between regional factors and the health of the residents
Q3	Use of the system can effectively uncover patterns of disease association
Q4	The system can effectively expand the user's medical knowledge
Q5	The system can give a reasonable prediction of the patient's potential high risk of disease

A. VISUAL DESIGN
All experts agreed that the system uses various types of visual charts and patterns to intuitively present abstract disease association patterns and the overall health status of the population. ''The presentation of disease associations in the form of network structure diagrams is very visual and intuitive and makes it easier to identify some interesting patterns of association compared with traditional textual and numerical interface-based systems''. Experts commented, ''It is very inspiring for the medical work we are currently doing''.

B. USER INTERACTIONS
The interactivity of the system was confirmed by all experts. One expert was impressed by the search function, which allows him to search for a given disease and obtain a simultaneous update of almost all views in the interface, thus keeping track of all aspects of the disease in detail. In addition, experts have suggested integrating the functions of the three system interfaces and creating cross-interface interactions to make the system functions more cohesive.

C. USABILITY
All experts agreed that our work had rich research and application value. ''The work to explore the relationship between urban area factors and the health of the population is very novel, allowing users to quickly understand the geographical and environmental characteristics of the city, which will help them to build awareness of the regional environment and diet-induced high-risk diseases'', experts commented. In addition, experts suggested that we could create an interactive guide with informative animations so that nonexpert users could become familiar with the system's functions and usage more quickly.

VIII. DISCUSSION
Taking the electronic medical record data and regional factors such as restaurant, industry, economy and climate data of a city as examples, this paper realizes visualization to explore the association pattern between various regional factors and residents' health in this city. Due to the many research areas involved, much of the work will be further refined in future studies.

In conducting the mining work on regional dining habits and residents' health, the data used is information related to online restaurants. This information plays an important role in mining the eating habits of people such as foreign tourists and general workers; however it is slightly insufficient for mining the eating habits of permanent residents, and the subjective nature of the review data can adversely affect the analysis results. In future work, we will try to obtain local food purchase data so that the system can more objectively and comprehensively explore the eating habits of each population.

The interviewed experts and volunteers involved in the work of this paper and all personal patient data used were treated confidentially. Therefore, in our disease prediction work, we are unable to interview the predicted patients to verify the accuracy of the prediction results. However, by predicting the course of disease, patients can be reminded to correct their bad habits in time to avoid the migration and deterioration of the disease, thus improving their health. Further work in the future will attempt to collaborate with hospitals in the city to obtain more detailed data on patient signs while using deep learning methods to further refine the disease prediction model.

IX. CONCLUSION AND FUTURE WORKS
In this paper, we propose a general analysis process to explore the association patterns between regional factors and urban residents' health, and we design and implement an interactive analysis system with multiple visualization interfaces. The Apriori association algorithm, one-mode projection algorithm for bipartite networks and collaborative filtering recommendation algorithm are used to explore the association of diseases, predict the potential high-risk diseases of patients, and complete disease popularization functions. A case study from electronic medical record data of a prefecture-level city was conducted to initially explore the association patterns of regional factors on the health of the city's residents, demonstrating the effectiveness of the generic analysis process proposed in this paper.

In future research, this work will be improved in the following respects: (1) The system and related algorithms will be improved based on the evaluation feedback results. (2) The China Health and Nutrition Survey (CHNS) data, urban travel data, and health care consumption data will be combined to further extend the relationship between regional factors and residents' health. (3) Deep learning algorithms will be introduced to explore the association patterns between regional factors and residents' health in a more comprehensive, detailed and accurate way.

ACKNOWLEDGMENT
The authors wish to thank anonymous reviewers for valuable and insightful comments. This work was supported by

National Natural Science Foundation of China Project (Grant No.61802320, No.61872304) and the PhD Research Foundation of Southwest University of Science and Technology (Grant No.19zx7144). The authors sincerely thank the Shandong Public Data Open Network (DATA.SD.GOV.CN) for supplying the data.

REFERENCES

[1] G. Markowitz and D. Porter, "Health, civilization and the state: A history of public health from ancient to modern times," *J. Public Health Policy*, vol. 21, no. 4, p. 488, 2000, doi: 10.4324/9780203980576.

[2] H. Li and L. Huang, "Health, education, and economic growth in China: Empirical findings and implications," *China Econ. Rev.*, vol. 20, no. 3, pp. 374–387, Sep. 2009, doi: 10.1016/j.chieco.2008.05.001.

[3] J. Strauss and D. Thomas, "Health, nutrition and economic development," *J. Econ. Literature*, vol. 36, no. 2, pp. 766–817, 1998, doi: 10.2307/1183456.

[4] J. Yang, Y. Li, Q. Liu, L. Li, A. Feng, T. Wang, S. Zheng, A. Xu, and J. Lyu, "Brief introduction of medical database and data mining technology in big data era," *J. Evidence-Based Med.*, vol. 13, no. 1, pp. 57–69, Feb. 2020, doi: 10.1111/jebm.12373.

[5] B. C. Kwon, V. Anand, K. A. Severson, S. Ghosh, Z. Sun, B. I. Frohnert, M. Lundgren, and K. Ng, "DPVis: Visual analytics with hidden Markov models for disease progression pathways," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 9, pp. 3685–3700, Sep. 2021, doi: 10.1109/TVCG.2020.2985689.

[6] Y. Zhang, K. Chanana, and C. Dunne, "IDMVis: Temporal event sequence visualization for type 1 diabetes treatment decision support," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 512–522, Jan. 2019, doi: 10.1109/TVCG.2018.2865076.

[7] R. Estruch, E. Ros, J. Salas-Salvadó, M.-I. Covas, D. Corella, F. Arós, E. Gómez-Gracia, V. Ruiz-Gutiérrez, M. Fiol, J. Lapetra, and R. M. Lamuela-Raventos, "Primary prevention of cardiovascular disease with a Mediterranean diet," *New England J. Med.*, vol. 368, no. 14, pp. 1279–1290, 2013, doi: 10.1056/NEJMoa1200303.

[8] S. Jadsri, P. Singhasivanon, J. Kaewkungwal, R. Sithiprasasna, S. Siriruttanapruk, and S. Konchom, "Spatio-temporal effects of estimated pollutants released from an industrial estate on the occurrence of respiratory disease in maptaphut municipality, Thailand," *Int. J. Health Geographics*, vol. 5, no. 1, pp. 1–13, 2006, doi: 10.1186/1476-072X-5-48.

[9] F. J. Kelly and J. C. Fussell, "Air pollution and airway disease," *Clin. Experim. Allergy*, vol. 41, no. 8, pp. 1059–1071, Aug. 2011, doi: 10.1111/j.1365-2222.2011.03776.x.

[10] W. Sun, Z. Cai, F. Liu, S. Fang, and G. Wang, "A survey of data mining technology on electronic medical records," in *Proc. IEEE 19th Int. Conf. e-Health Netw., Appl. Services (Healthcom)*, Oct. 2017, pp. 1–6, doi: 10.1109/HealthCom.2017.8210774.

[11] M. Glueck, M. P. Naeini, F. Doshi-Velez, F. Chevalier, A. Khan, D. Wigdor, and M. Brudno, "PhenoLines: Phenotype comparison visualizations for disease subtyping via topic models," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 371–381, Jan. 2018, doi: 10.1109/TVCG.2017.2745118.

[12] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo, "RetainVis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 299–309, Jan. 2019, doi: 10.1109/TVCG.2018.2865027.

[13] A. B. Hill, "The environment and disease: Association or causation?" *J. Roy. Soc. Med.*, vol. 108, no. 1, pp. 32–37, Jan. 2015, doi: 10.1177/0141076814562718.

[14] S. B. Abdullahi, K. Muangchoo, A. B. Abubakar, A. H. Ibrahim, and K. O. Aremu, "Data-driven AI-based parameters tuning using grid partition algorithm for predicting climatic effect on epidemic diseases," *IEEE Access*, vol. 9, pp. 55388–55412, 2021, doi: 10.1109/ACCESS.2021.3068215.

[15] A. A. Bui and W. Hsu, "Medical data visualization: Toward integrated clinical workstations," in *Medical Imaging Informatics*. Boston, MA, USA: Springer, 2010, pp. 139–193, doi: 10.1007/978-1-4419-0385-3_4.

[16] M. H. Loorak, C. Perin, N. Kamal, M. Hill, and S. Carpendale, "TimeSpan: Using visualization to explore temporal multi-dimensional data of stroke patients," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 409–418, Jan. 2016, doi: 10.1109/TVCG.2015.2467325.

[17] J. Bernard, D. Sessler, J. Kohlhammer, and R. A. Ruddle, "Using dashboard networks to visualize multiple patient histories: A design study on post-operative prostate cancer," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 3, pp. 1615–1628, Mar. 2019, doi: 10.1109/TVCG.2018.2803829.

[18] M. Elshehaly, R. Randell, M. Brehmer, L. McVey, N. Alvarado, C. P. Gale, and R. A. Ruddle, "QualDash: Adaptable generation of visualisation dashboards for healthcare quality improvement," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 689–699, Feb. 2021, doi: 10.1109/TVCG.2020.3030424.

[19] Z. Zhang, B. Wang, F. Ahmed, I. V. Ramakrishnan, R. Zhao, A. Viccellio, and K. Mueller, "The five Ws for information visualization with application to healthcare informatics," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 11, pp. 1895–1910, Jun. 2013, doi: 10.1109/TVCG.2013.89.

[20] W. C. Willett, J. P. Koplan, R. Nugent, C. Dusenbury, P. Puska, and T. A. Gaziano, "Prevention of chronic disease by means of diet and lifestyle changes," *Disease Control Priorities Developing Countries*, 2nd ed. Washington, DC, USA: The World Bank, 2006.

[21] K. B. Michels and M. B. Schulze, "Can dietary patterns help us detect diet–disease associations?" *Nutrition Res. Rev.*, vol. 18, no. 2, pp. 241–248, Dec. 2005, doi: 10.1079/NRR2005107.

[22] F. Violi, D. Pastori, P. Pignatelli, and R. Carnevale, "Nutrition, thrombosis, and cardiovascular disease," *Circulat. Res.*, vol. 126, no. 10, pp. 1415–1442, May 2020, doi: 10.1161/CIRCRESAHA.120.315892.

[23] Y. Yuexin, *China Food Composition*. Beijing, China: Peking Univ. Medical Press, 2005.

[24] W. Xiaolong and H. Sun, "A method of radiation range estimation in public safety issue caused by point source pollution," in *Proc. Innov. Theories Methods Risk Anal. Crisis Response*, 2012, pp. 680–683.

[25] Z. Shuzhen and S. Jiong, *Urban Climatology*. Beijing, China: China Meteorological Press, 1994.

[26] A. M. Khedr, Z. A. Aghbari, A. A. Ali, and M. Eljamil, "An efficient association rule mining from distributed medical databases for predicting heart diseases," *IEEE Access*, vol. 9, pp. 15320–15333, 2021, doi: 10.1109/ACCESS.2021.3052799.

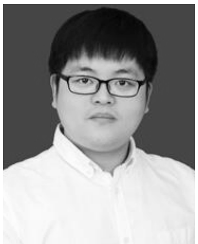
[27] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang, "Bipartite network projection and personal recommendation," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, no. 4, Oct. 2007, Art. no. 046115, doi: 10.1103/PhysRevE.76.046115.

[28] S. Wang, Y. Zhang, and W. U. Yadong, "Survey on network topology visualization," *Chin. J. Netw. Inf. Secur.*, vol. 4, no. 2, p. 1, 2018.

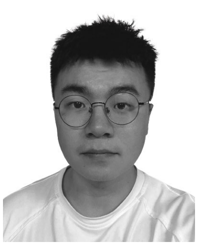
[29] M. A. Riva, A. Lafranconi, M. I. D'orso, and G. Cesana, "Lead poisoning: historical aspects of a paradigmatic 'occupational and environmental disease,"' *Saf. Health Work*, vol. 3, no. 1, pp. 11–16, 2012, doi: 10.5491/SHAW.2012.3.1.11.

[30] J. C. Antunes, J. G. L. Frias, A. C. Micaelo, and P. Sobral, "Resin pellets from beaches of the Portuguese coast and adsorbed persistent organic pollutants," *Estuarine, Coastal Shelf Sci.*, vol. 130, pp. 62–69, Sep. 2013, doi: 10.1016/j.ecss.2013.06.016.

TANGYUAN ZOU received the B.S. degree from the Southwest University of Science and Technology, Mianyang, China, in 2019, where he is currently pursuing the M.S. degree with the School of Computer Science. His major research interest includes information visualization.



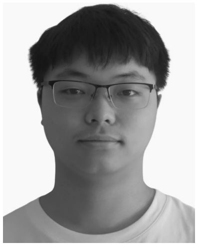
SONG WANG received the Ph.D. degree from the Institute of Electronic Engineering, Academy of Engineering Physics, China, in 2019. He is currently the Deputy Director of the Software Engineering Department, School of Computer Science and Technology, Southwest University of Science and Technology. He presided over one National Natural Science Youth Fund Project, one Sichuan Miaozi Project, and one Doctoral Fund Project. In the past three years, he has published more than ten academic papers in foreign journals and conferences, six of which were indexed by SCI/EI, applied for six invention patents, and obtained five software copyrights. His main research interests include visualization and visual analysis, research content includes flow field visualization, medical visualization, network security visualization, and urban computing visualization. He is a member of IEEE CS, CCF, and CSIG.



YIFAN ZHANG received the B.S. degree from the Chengdu University of Information Technology, Chengdu, China, in 2019. He is currently pursuing the M.S. degree with the School of Computer Science, Southwest University of Science and Technology, Mianyang, China. His major research interest includes information visualization.

YING ZHONG received the B.S. degree from the Southwest University of Science and Technology, Mianyang, China, in 2020, where she is currently pursuing the M.S. degree with the School of Computer Science. Her major research interest includes information visualization.

SHIJIE CHEN received the B.S. degree from the Southwest University of Science and Technology, Mianyang, China, in 2021, where he is currently pursuing the M.S. degree with the School of Computer Science. His major research interest includes information visualization.



CHAO FU is currently pursuing the B.S. degree with the School of Computer Science, Southwest University of Science and Technology, Mianyang, China. His major research interest includes information visualization.