




**Please cite the Published Version**

Yan, Lu, Huang, Weihong , Wang, Liming , Feng, Song, Peng, Yonghong  and Peng, Jie (2021) Data-enabled digestive medicine: a new big data analytics platform. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18 (3). pp. 922-931. ISSN 1545-5963

**DOI:** <https://doi.org/10.1109/TCBB.2019.2951555>

**Publisher:** IEEE

**Version:** Published Version

**Downloaded from:** <https://e-space.mmu.ac.uk/635036/>

**Usage rights:**  [Creative Commons: Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)

**Additional Information:** This is an open access article which first appeared in IEEE/ACM Transactions on Computational Biology and Bioinformatics

**Enquiries:**

If you have questions about this document, contact [openresearch@mmu.ac.uk](mailto:openresearch@mmu.ac.uk). Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

***IEEE Xplore*** ®

**Notice to Reader**

“Data-Enabled Digestive Medicine: A New Big Data Analytics Platform”

By Lu Yan, Weihong Huang, Liming Wang, Song Feng, Yonghong Peng, and Jie Peng

published in *IEEE/ACM Transactions on Biology and Bioinformatics*

vol. 18, no. 3, pp. 922-931, May/June 2021

Digital Object Identifier: 10.1109/TCBB.2019.2951555.

The corresponding author of this article is Jie Peng  
(e-mail: pengjie2014@csu.edu.cn.)

We regret any inconvenience this may have caused.

Srinivas Aluru

Editor-in-Chief

*IEEE/ACM Transactions on Biology and Bioinformatics*

# Data-Enabled Digestive Medicine: A New Big Data Analytics Platform

Lu Yan, Weihong Huang<sup>1</sup>, Liming Wang<sup>2</sup>, Song Feng, Yonghong Peng<sup>3</sup>, and Jie Peng

**Abstract**—This paper presents a big data analytics platform for clinical research and practice in the Gastroenterology Department of Xiangya Hospital at Central South University in China. This platform features a comprehensive and systematic support of big data in digestive medicine including geneneral health management, clinical gastroenterology practice, and related genomics research, which is proven to be helpful in real world clinical practices. A typical use case of integrated analysis based on electronic medical records and colonoscopy data was presented and discussed, the analaystic report on risk factors of colorectal diseases shows a reasonable recommendation about the age when people should start to screen the colorectal cancer, which could be very useful to individual and group health management for the general population in China.

**Index Terms**—Digestive medicine, big data analystic platform, gastroenterology research, colorectal cancer screen

## 1 INTRODUCTION

THE wave of big data has swept across the world, and data-driven healthcare has become an important field of big data application. The application of big data technology in healthcare is expected to enable better health and medical services, as well as to discover new medical knowledge and to promote the progress of clinical practices continuously [1].

Big data for healthcare has been adopted as a national strategy by many countries. For example, the British government announced the launch of the 100,000 genome project (100,000 Genomes Project) for cancer and rare diseases in 2012. The United States launched the project of ‘big data research and development initiative’ for purpose of using big data in other related fields for scientific exploration, discovery and biomedical research in 2012, and the NIH ‘Big Data to Knowledge initiative’ (BD2K) plan in 2014. Compared with some foreign developed countries, China’s big data development started lately. As a developing country, China considers big data as one of the important future strategies for national developments [2]. In 2015, China issued an action plan to promot the development of big data in different industries including healthcare [3]. It is proposed to construct medical and health service big data including electronic

health records and electronic medical records, build a medical and health management and service big data application system covering public health, medical services, medical security, drug supply, family planning and comprehensive management business, and carry out innovative application research on medical and health big data [4]. In most cases healthcare big data projects are initiated by universities and supported by large hospitals [5]. In January 2014, the Central South University launched the “Xiangya clinical big data system project”, it includes 101 projects in the first phrase which cover more than 40 clinical specialities. This project is the first large-scale and systematic exploration and application of big data in clinical medicine in China. Through clinical big data mining and analysis, the project establishes an advanced cutting-edge medical big data framework towards intelligent medicine, precision medicine, hospital fine management, clinical research, translational medicine and public health support for decision-making by local health authorities [6].

Digestive system disease is one of the most common diseases supported by the big data framework. It features high diversity and complexity, and in many cases interrelated with other disease, which could seriously affect the quality of life of patients. The data of China health and family planning statistical yearbook in 2013 showed that the ratio of digestive system diseases to all diseases was 10.12 to 11.28 percent [7]. Therefore, the health care big data platform is suitable for complex analysis of clinical data, the auxiliary of the digestive system disease early warning and prevention, early detection and treated, which has important clinical significance and value and is contributing to improve the level of the health of the people. However, the relevance studies about the big data of digestive system diseases are rarely reported. Under the framework of the “Xiangya clinical big data” and supported by the Gastroenterology Department of Xiangya Hospital, the gastroenterology big data platform for clinical research and practice is initiated. This platform takes the advantage of a comprehensive view of patients’ data of

- L. Yan and J. Peng are with the Department of Gastroenterology, Xiangya Hospital, Central South University, Changsha 410008, China. E-mail: {yanluxy, pengjie2014}@csu.edu.cn.
- W. Huang is with the Mobile Health Ministry of Education - China Mobile Joint Laboratory, Xiangya Hospital, Central South University, Changsha 410008, China. E-mail: whuang@csu.edu.cn.
- L. Wang is with Bitvalue Technology (Hunan) Co. Ltd. Xiangjiang Road, Changsha 410082, China. E-mail: yinhu0415@126.com.
- S. Feng is with the Network Information Center, Xiangya Hospital, Central South University, Changsha 410008, China. E-mail: fs205@sina.com.
- Y. Peng is with the Faculty of Computer Science, University of Sunderland, Sunderland SR6 0DD, UK. E-mail: yonghong.peng@sunderland.ac.uk.

Manuscript received 31 Jan. 2019; revised 1 July 2019; accepted 30 Oct. 2019.  
Date of publication 6 Nov. 2019; date of current version 3 June 2021.  
(Corresponding author: Weihong Huang.)  
Digital Object Identifier no. 10.1109/TCBB.2019.2951555

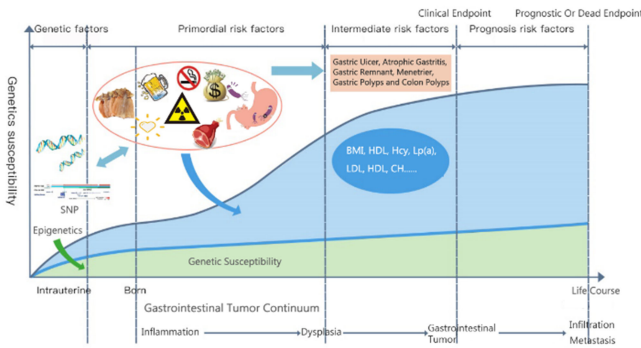


Fig. 1. The gastrointestinal tumor continuum (GTC) concept model.

personal health management, clinical medicine, genomics overall. By helping the clinicians in analysis and problem solving, it aims to promote the research achievements and more precise clinical practices in gastroenterology for sustainable development of the department and the hospital.

Furthermore, this platform has been gradually opened to other clinical departments to support different clinical research work in different diseases. We are trying to build up the health care big data platform which is suitable for diseases of all systems within our whole hospital. And at present, multi-center clinical research is of great significance for realizing multi-center and multi-discipline extensive collaborative research on the same clinical problem. Our plan is to build this big data platform into a multi-center clinical big data application platform, which will cover the most of the large and medium-sized hospitals in the whole Hunan province.

The rest of this paper is organized as follows: Section 2 introduces the background information of clinical research in gastroenterology, Section 3 illustrates the technical framework of the big data analytics platform in gastroenterology, Section 4 presents a case study of data-enabled gastroenterological research scenario, Section 5 discusses the potential and limitations of the work and Section 6 concludes the paper.

## 2 CONCEPTUAL MODEL AND CHALLENGES OF DIGESTIVE MEDICINE IN BIG DATA ERA

This section introduces the conceptual model of digestive diseases, data scale and the major challenges faced by data-enabled research in digestive medicine.

### 2.1 The Conceptual Model of Digestive Diseases

The process of any disease from occurrence and development to a certain outcome (complications, death, etc.) is the result of the gradual accumulation of risk factors. Early detection and prevention of the diseases or delay the progress of diseases through health care management is proven to be effective [8].

To guide the theoretical research and clinical practice, a conceptual model of digestive medicine is proposed. This model covers the lifecycle of continuous time dimension on all disease-causing genes and the environment factor and their sequence on the chain, and especially emphasized on the chain 'gene-environment' gastrointestinal lesions are regulated by a complex interaction always occurrence,

development and outcome process, make the digestive tract health management throughout the life course. In order to highlight the importance of 'early life course health management', we divided the risk factors into 'genetic risk factors', 'initial risk factors', 'intermediate risk factors' and 'return risk factors' according to the occurred sequence of risk factors.

Digestive system diseases have a wide variety of types and an increasing incidence, while precancerous lesions and early cancers of the digestive tract are major diseases of the digestive system [9]. An example of the gastrointestinal tumor continuum (GTC) could show how the conceptual model of digestive diseases works (Fig. 1).

In the conceptual model of GTC: (1) genetic risk factors (GRF): pathogenic genes and regulatory factors of gastrointestinal tumors and their main risk factors (e.g., gastric ulcer, atrophic gastritis, gastrointestinal polyps, Menetrier's disease, etc.); (2) "primordial risk factors (PRF)": early interventional risk factors such as life behavior, dietary habits, psychological spirit, socio-economic, environmental, helicobacter pylori, etc.; (3) "intermediate risk factors (IRF)": risk factors that promote the occurrence or occurrence of gastrointestinal tumors (e.g., gastric ulcer, atrophic gastritis, gastrointestinal polyps, Menetrier's disease, gastric stump, etc.); (4) Prognosis risk factors (ProRF): risk factors in the prognosis. This classification aims to focus on the whole life-cycle, and to guide disease early detection, risk assessment and personalized intervention for key links in the chain of events in the digestive tract. When the health/disease detection option is selected and determined with conclusive evidence, the best disease screening indicators and predictors of health risk assessment are to build a health/disease prediction model and corresponding evaluation method, and a consequent intervention method is provided based on the theory of causal inference results which have certain causal relationship between risk factors and intervention and effect of food or medicine.

The core value of this conceptual model of digestive tract diseases is to the best utilization of health and medical data, which could be useful to precise supply of healthcare resources, reduced medical costs, improved quality and efficiency of care services, and could lead to revolutionary changes to the health care model of digestive system.

### 2.2 The Accumulate Data of Digestive Diseases

In the past decade, China has benefit greatly from the development of healthcare information systems in hospitals and administrative authorities around the country [10]. Electronic medical records (EMR) and electronic health records (EHR) normally contain the information of patient profile, registration, medical images and reports, examination reports of biological samples, regular health examine reports, community residents' health records, infectious disease register, medical insurance and demographics, etc. For example, the accumulated dataset of the Department of Gastroenterology of Xiangya Hospital between 2010 and 2018 is shown as follows (Table 1).

In terms of the value of the data, it is quite obvious to make the best use of the dataset to provide data-enabled digestive medicine, and offer best health management advices for patients and the general population.

TABLE 1  
Accumulated Dataset of the Department of Gastroenterology  
of Xiangya Hospital (2010-2018)

| Data Type                  | Volume                 |
|----------------------------|------------------------|
| Out-patient                | 1,032,995 person-times |
| Hospitalization            | 32,864 person-times    |
| Health examination         | 586,938 person-times   |
| Laboratory Test            | 7,392,364 items        |
| Inspection Report          | 474118 items           |
| Medication record          | 1,605,400 items        |
| Surgery Record             | 16,663 items           |
| Medical Text               | 1,157,309 items        |
| Gastrointestinal Endoscopy | 1.58 TB                |
| CT                         | 16.46 TB               |
| MRI                        | 4.75 TB                |
| X-ray                      | 0.57 TB                |

### 2.3 Requirements for Data-Enabled Digestive Medicine Research Support

#### 2.3.1 Difficulties in Traditional Clinical Trial and Evidence-Based Digestive Medicine

In traditional clinical trials in digestive medicine in China, experiments are hard to manage due to the complexity in the lengthy procedure and quantitative data collection and analysis after trials, which normally incur high cost in time and human resources.

Evidence-based medicine emphasizes the combination of clinical expertise, patient values and desires, and the best external research evidence currently available to make treatment measures for patients [11]. However, there are some problems such as the lack of high-quality evaluation of medical information technology and the gap between the research theory and practice of evidence-based medical informatics [12]. For example, in 2015, only 13 percent of the Japanese gastroenterological endoscopy society guidelines for ESD and EMR for early gastric cancer had scientific evidence (recommended), i.e., class B recommendation intensity, while no recommendation intensity was class A (highly recommended with high level of scientific evidence) [13]. This suggests that evidence-based medicine could only be more useful for research and practice when data is available for analysis and helpful to clinical decision making.

#### 2.3.2 Data Issues in Clinical Research in Digestive Medicine

Doctors in big hospitals in China are very busy with daily clinical work, and most of their scientific research work is still in the stage of manual arrangement without systematic scientific research support platform. Lab research assistants' daily work include obtaining clinical data from production systems such as hospital information systems (HIS) and EMR by exporting Excel files, and then use general statistical software for analysis, which means the process is low efficient and with high risks of manual data entry errors. On top of that, integration of heterogeneous data sources is normally a "mission impossible" without a proper information platform support.

In addition, most of the digestive tract diseases have a long treatment cycle, which requires long-term follow-up of

patients' treatment, inquiry of historical medical records and comparison of historical data. Manual collation could not meet the needs of scientific research management. It leads to difficulties in medical record searching, filing, data collection and individualized treatment.

## 3 BIG DATA ANALYTICS PLATFORM FOR DIGESTIVE MEDICINE

Based on the real-world requirements in gastroenterological research and practice, a digestive medicine big data analytic platform was built in Xiangya Hospital with full support from the hospital information integration platform and clinical data repository (CDR), which is a safe and sufficient solution using the secondary data processing approach [14].

In order to solve the pre-analysis problems of information system isolation, data complexity, duplicated information, difficulty in data extraction, etc., all data preparation work is done in the CDR with standardized data model in HL7 V3 reference information model (RIM) and refined message information models (RMIM). Classic techniques such as ELT is used to extract, transform and load the data from the different hospital information systems into a uniform data center, without interfering with the operation of the hospital's existing production system. After that, data is processed according to relevant specifications oriented to gastroenterological research. The big data analytic platform then supports clinical research on top of the existing big data from heterogeneous sources including offering the research staff for extra research data input in addition to the information of the EMR with concise and standard two-dimensional form, which could be managed easily on the platform for ultimate analysis of data.

### 3.1 Architecture of the Digestive Medicine Analytic Platform

The technical framework of digestive medicine scientific research platform was based on the data center operating system (DCOS). DCOS provides the distributed scheduling and resource coordination functions for the whole data center, and enabled the data center resilient scalable software stack. Based on the DCOS, massive medical and health data could be stored and retrieved safely and reliably by integrating a series of big data frameworks, such as Hadoop, Hbase, ElasticSearch, Hive, and so on. Data governance, fusion and machine learning algorithm support for healthcare data analysis are enabled, and some computing frameworks and techniques, such as Spark and R are integrated to ensure the unified management, scheduling and monitoring of resources.

The functional structure of digestive medicine analytic platform is organized in four levels: application layer, data mart layer, clinical data center layer and business system. The specific architecture is shown in Fig. 2.

The business system layer is the data source of the research platform, including HIS, CIS, NIS, EMR, PACS and LIS. The generated data includes various medical information of patients, such as medical history information, clinical diagnosis information, inspection and examination information, nursing information, medication information and surgical information, etc.

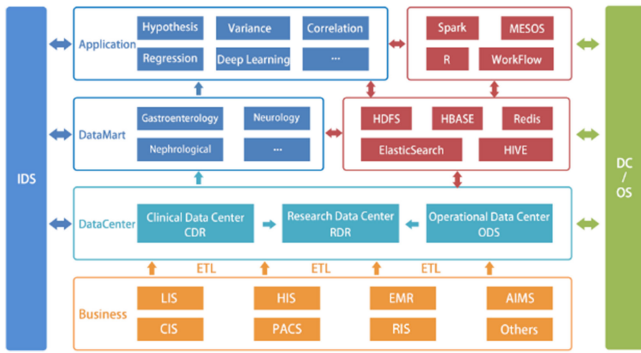


Fig. 2. Architecture of digestive medicine analytic platform.

The data generated by the business system is collected, filtered, transformed, and structured and stored in the clinical data center layer. The clinical data center layer is used to process, store and manage all kinds of data of the whole platform, including all the important clinical data of the patient, and establish the EMPI (Enterprise Master Patient Index) to form a hierarchical data storage structure with the EMPI as a clue. The clinical data center realizes the integration and centralized display of all clinical diagnosis and treatment data, and the data can be easily processed and arranged to provide support information for medical decision-making and scientific research activities.

Based on the clinical data center, a special disease data mart was constructed. Specialized disease data marts preset the characteristics of various single disease indicators of digestive medicine, such as clinical manifestations, complications, infection indicators, care intake, test results, drugs, test results, bacterial identification culture, drug susceptibility results, etc. These indicators can be maintained by the researchers themselves, such as an increase in indicators based on literature results. The special disease data mart automatically extracts the required data from the clinical data center according to the predefined characteristics of various indicators, and conducts various scientific research and analysis on the specific diseases. Specialized disease data marts are more targeted and meet the specific needs of research users in terms of analysis, content, performance and easily for use.

The application layer uses the data of the special disease data market layer as the material to realize scientific research and application. The application layer provides applications such as medical record retrieval, specialist view, scientific patient data grouping and scientific research follow-up, and provides a wealth of tools for data exploration, visualization, and analysis mining. Researchers can complete high-end, personalized scientific research and analysis tasks by clicking or dragging.

### 3.2 Featured Functions of the Digestive Medicine Analytic Platform

#### 3.2.1 Multi-Center Support

The digestive medicine scientific research platform supports multi-center data collection, data fusion and scientific research collaboration, which means a single platform could be used among numbers of hospitals and community health centers if necessary for broader coverage in digestive medicine research.

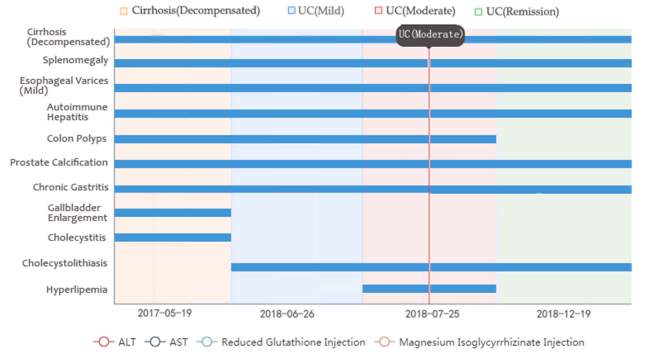


Fig. 3. View of the gastroenterology disease evolutionary and the relationship between drugs and laboratory test.

By collecting multi-center medical health data and fusing data, it enriches and improves patients' medical health data, and provides richer and more reliable data support for patients' diagnosis and health management. Through multi-center joint cooperation with data safety assurance mechanisms, data completeness and integrity could be achieved for better research efficiency and better quality research outcomes.

#### 3.2.2 Research Case Retrieval

The scientific medical record retrieval system has changed the method of retrieving scientific medical records from different business systems in the past. The system is based on clinical data centers, and can search for medical records which meet the indicators from the disease data marts according to single or multiple combined conditions. The search results include information such as outpatient diagnosis, hospitalization diagnosis, inspection details, inspection results, inspection details, and inspection results.

#### 3.2.3 Special Disease Diagnosis View

The special disease diagnosis view shows all the medical data generated during the entire process from the initial diagnosis to hospital discharge of a specialized case. With the time axis and the presentation dimension of the type of medical information, the information generated by the case during the diagnosis and treatment is comprehensively and accurately displayed. The special disease diagnosis view is a more refined application of the unified clinical view. It can deeply display the scientific research value of the medical information by directing at the specialist scientific cases. Fig. 3 shows the special disease diagnosis view of a patient with digestive disease. The upper part of this picture shows the evolutionary relationship between the patient's primary disease and complications. The lower part of this picture shows the value of the patient's medication and test indicators on the timeline, and monitor the patient's corresponding life index at different medications.

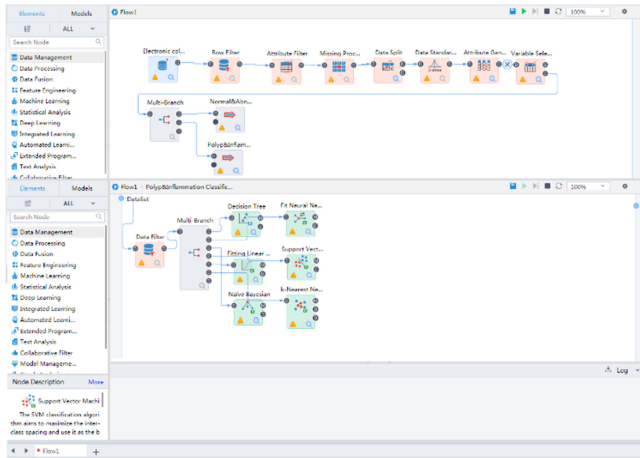


Fig. 4. Design of colonoscopy population analysis process based on self-service data analysis and mining tools.

Upper part of the figure shows the evolutionary relationship between a patient's main digestive tract disease and its comorbidities. Lower part of the figure shows the trend of ALT and AST over time after treatment with Reduced Glutathione Injection and Magnesium Isoglycyrrhizinate Injection in a drug-induced hepatitis patient.

### 3.2.4 Self-Service Data Analysis With Machine Learning Algorithms

The machine learning module enabled by the analytic platform is a refinement and summary of classic machine learning models, and trained with various algorithms [15]. The platform applies those common machine learning algorithms in healthcare data domain and encapsulates it into a series of machine learning service options, which frees users from the problems of algorithm implementation, simplifies the research and analysis process, and improves scientific research efficiency. Fig. 4 shows an example for how to use machine learning for self-service analysis.

The upper part of Fig. 4 illustrates the overall analysis process for the colonoscopy population. The process also includes two sub-processes for data analysis modeling, which are predictions of normal and abnormal outcomes for all populations and predictions of inflammation and polyps in abnormal outcomes. The lower part of Fig. 4 shows a sub-process for modeling normal and abnormal outcome predictions for general public.

## 4 CASE STUDY

This section introduces a real-world digestive medicine research case using the analytic platform presented above.

### 4.1 Study Introductions

The intestinal canal is the longest tube in the digestive system, and most of the intestinal diseases are located in the lower digestive tract. With the changes in the working methods, living habits and dietary structure of the people in our country, the incidence of lower digestive tract diseases has increased, and the structure of intestinal diseases has also changed accordingly. Since 2010, cancer has become the leading cause of death in China, and the incidence of colorectal cancer is

still increasing, which is one of the most common digestive tract tumors [16]. Therefore, understanding the distribution and risk factors of intestinal diseases in the population is crucial for clinicians, endoscopic surgeons and medical examiners.

## 4.2 Experimental Data and Methods

### 4.2.1 Experimental Data

This study included 4,150 patients who underwent physical examination at the Xiangya Hospital of Central South University and the Hunan Provincial People's Hospital from June 2011 to December 2018 and completed an electronic colonoscopy. The characteristics of colorectal lesions detected by colonoscopy in the physical examination population were analyzed. There were 1,585 patients with negative colonoscopy, accounting for 38.19 percent, and 2,565 patients with positive colorectal disease, accounting for 61.81 percent. The total detection rate was higher than other regions. In terms of disease distribution, non-specific colorectal inflammation has a high detection rate (43.63 percent) and is higher than other regions, which may be related to the geographical difference of distribution of colorectal diseases, including the influence of living habits and economic development level.

### 4.2.2 Data Pre-Processing

The first step is to analyze the patient's pathology and microscopy results. If the result was diagnosed as cancer, the patient was excluded; if the result is diagnosed as a "polyp", the patient setting data is labeled as a polyp; if the result is diagnosed as "inflammation", The patient setting data is labeled as inflammation; if the results are diagnosed with both "polyps" and "inflammation", the patient setting data is labeled as a polyp; the remaining data labels are set to normal.

The second step excludes patient records with a deletion rate greater than 60 percent in the samples, as well as a missing feature column of more than 60 percent.

The third step performs a normal distribution check on all features of the sample and finds that all features satisfy the normal distribution.

The fourth step filled in missing values for all features of the sample. For the feature blood pressure and BMI, the samples were first grouped according to age and gender, and the mean blood pressure and BMI of different ages and gender groups were calculated, then the missing blood pressure and BMI of the same age and gender were filled with the grouped average value; The blood pressure and BMI, which are still empty, are filled with the overall average. For other features, the overall average is used for filling.

### 4.2.3 Feature Selection by Stepwise Linear Regression Analysis

Stepwise Linear Regression Analysis is a method in Multiple Linear Regression (MLR), which is a classical regression analysis method for analyzing the linear relationship between a dependent variable and multiple independent variables. In order to eliminate the independent variables whose regression effects are not significant enough, combined with the stepwise regression method, the final generated prediction

TABLE 2  
The Description of the Data Used About the Features

| Feature | Missing (%) | Mean ± Std    |
|---------|-------------|---------------|
| Gender  | 0           | 0.71 ± 0.45   |
| Age     | 0           | 46.81 ± 9.68  |
| BMI     | 0           | 24.33 ± 3.59  |
| BS      | 0           | 5.4 ± 1.37    |
| TG      | 0           | 1.88 ± 1.55   |
| TC      | 0           | 5.15 ± 1.03   |
| LDL     | 0           | 3.19 ± 10.31  |
| HDL     | 0           | 1.45 ± 0.4    |
| CEA     | 0           | 3.21 ± 29.62  |
| SBP     | 0           | 121.16 ± 14.7 |
| DBP     | 0           | 77.43 ± 10.61 |
| FAT     | 0           | 0.67 ± 0.68   |

model is used in the subsequent verification analysis. In our study, the independent variables are the 12 characteristics of the sample, and the dependent variables are normal/abnormal colonoscopy results and polyps/inflammation of colonoscopy results. The features that contribute the most to the model are selected by stepwise linear regression analysis. And the description of the data used about the features are shown as follows (Table 2).

#### 4.2.4 Establishment of Prediction Model

Furthermore, we choose the features which were selected by the Stepwise Linear Regression model as the independent variables, and used nine different kind of machine learning algorithms [Fitting Linear Models(LR), k-Nearest Neighbour Classification(KNN), Random Forest(RF), Support Vector Machines(SVM), Bagging(BAG), Decision Tree(DT), Adaboost (AD), Naïve Bayesian(NB) and Fit Neural Networks (Nnet)] separately to establish a polyp/inflammation model which could predict normal/abnormal colonoscopy results and colonoscopy results.

#### 4.2.5 Model Validation

K-fold cross-validation is divided the initial sample into K parts (generally equal), each sub dataset is separately verified, and the remaining K-1 sub dataset is used as the training dataset, so K models are obtained. The average of the classification accuracy of the final verification set of the K models is used as the performance index of the classifier under this K-CV. In this paper, a 5-fold cross-validation is used to validate the model.

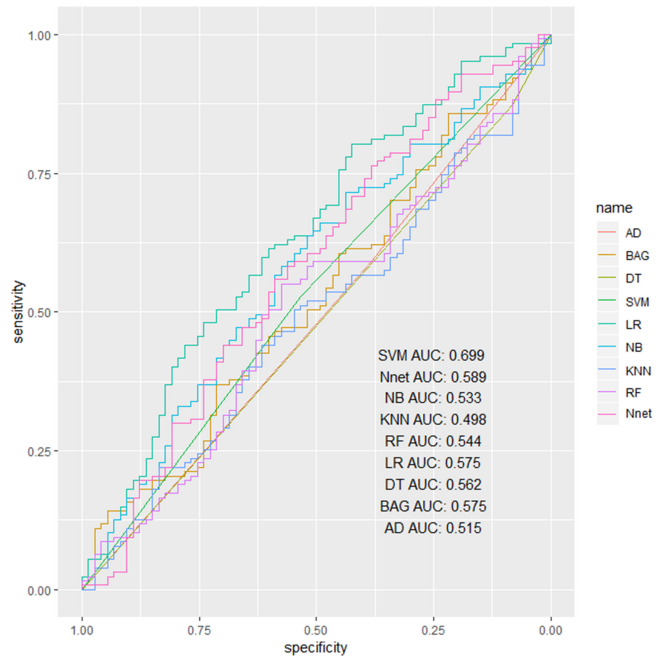


Fig. 6. The performance of colorectal diseases predicted model and comparison of nine prediction models' ROC curves.

### 4.3 Findings of Case Study

#### 4.3.1 Age and HDL are the Two Key Major Factors in Colorectal Diseases

To qualitatively measure the 12 characteristics (Sex, Age, BMI, Obesity, Blood Glucose, Triglycerides, Total Cholesterol, Low Density Lipoprotein, High Density Lipoprotein, Carcinoembryonic Antigen, Systolic Blood Pressure and Diastolic Blood Pressure) and knots involved in the sample Correlation of rectal diseases, the patients labeled with inflammation and polyps in the sample were classified as abnormal tags (2,565 patients), and a stepwise linear regression model was established. Fig. 5 shows the results of the stepwise regression equation and the correlation results. The correlation between each feature and the predicted result is given. The right graph shows the feature that contributes the most to the dependent variable after stepwise regression. We found that age, high-density lipoprotein, gender, BMI, and blood glucose contributed the most to the regression equation, with age and high-density lipoprotein being key factors influencing rectal disease. Further, age, high-density lipoprotein, gender, BMI, and blood glucose were used as independent variables, and the results of electronic colonoscopy were normal/abnormal as a dependent variable. Nine machine learning algorithms were used to establish a predictive model, and the effect of the predictive model was evaluated as shown in Fig. 6. We found that the prediction model based on the SVM algorithm works best and the AUC value reached 0.699.

#### 4.3.2 Colorectal Cancer Screening is Recommended Before the Age of 48

Age is one of the key factors affecting the occurrence of colorectal diseases. We further analyzed the age and found that there is an interesting connection between age and rectal disease. As shown in the figure (Fig. 7), the number of diseases occurring at the age of 48 has risen obviously.

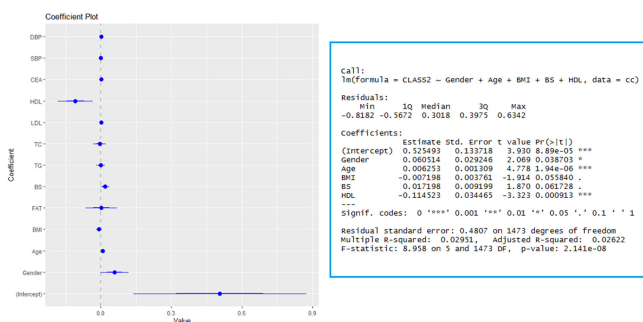


Fig. 5. Key factors related to colorectal diseases, through stepwise regression equation and the correlation analysis.



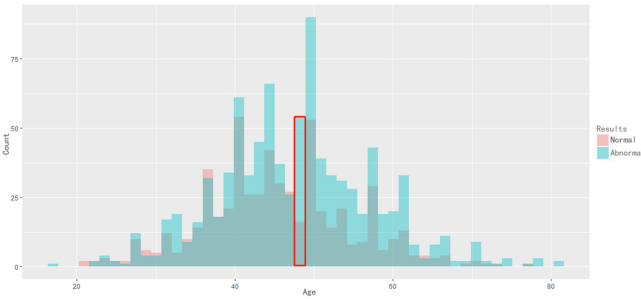


Fig. 7. Age distribution of results of colorectal cancer screening.

Furthermore, we divided the data by age. We set 45 to 50 years as the dividing line, and add 6 features to the data set. AGE-45 (two classification,  $\geq 45$  is 1,  $< 45$  is 0), AGE-46 (two classification,  $\geq 46$  is 1,  $< 46$  is 0), AGE-47 (two classification,  $\geq 47$  is 1,  $< 47$  is 0), AGE-48 (two classification,  $\geq 48$  is 1,  $< 48$  is 0), AGE-49 (two classification,  $\geq 49$  is 1,  $< 49$  is 0), AGE-50 (two classification,  $\geq 50$  is 1,  $< 50$  is 0). Then we used regression method to analyze the statistical significance of the above characteristics and dependent variables. The results are shown in the figure below (Fig. 8). The characteristic AGE-48 has the strongest statistical significance between the other independent variables and the dependent variable, and the p value is 0.000937.

The positive rate of intestinal diseases was higher in the healthy people with increasing age in our study as shown above. And the densities of distribution in the age 25, 30, 48, 50 were higher than others (Fig. 9). But the recommended starting age for colorectal cancer screening in China is 50 years old [17]. Combined with our research, we suggest that the age of colonoscopy for physical examination should be advanced to 48 years old.

### 4.3.3 Obesity is Closely Related to the Incidence of Colonic Polyps and Inflammation

To qualitatively measure the correlation between 12 features involved in the sample [gender, age, BMI, blood glucose, triglyceride (TG), total cholesterol (TC), low density lipoprotein (LDL), high density lipoprotein (HDL), carcino-embryonic antigen (CEA), systolic pressure (SBP) and diastolic blood pressure (DBP)] and colonic polyps and colitis, we selected the patients labeled as polyps and inflammation as training samples (2565 persons). And we established stepwise linear regression model (Fig. 10). In Fig. 10, the left part shows us the correlation between the features and the predicted

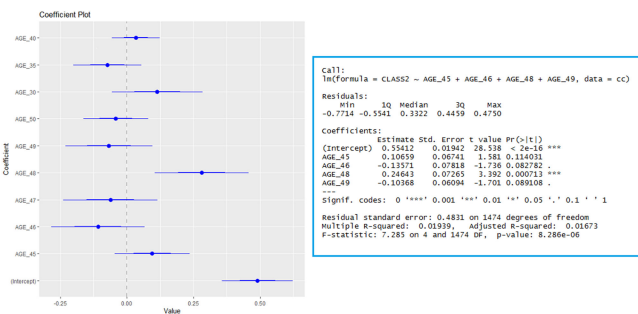


Fig. 8. The influence of different age groups on colorectal cancer, screening, through stepwise regression equation and the correlation analysis.

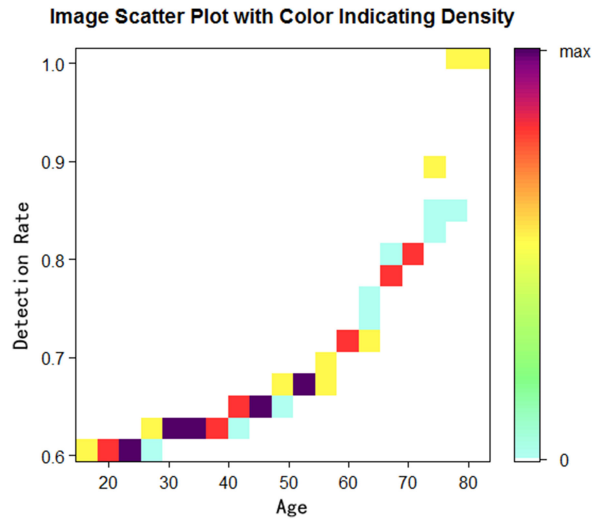


Fig. 9. Detection rate of colorectal cancer in different age groups and the densities of distribution of colorectal cancer patients.

results, and the right part is the feature selected after stepwise regression that contributed the most to the dependent variable. We found that gender, age, obesity contribute the most to the regression equation, and obesity was closely related to the incidence of colonic polyps and colitis.

Furthermore, we used gender, age and obesity as independent variables, and used the results of electronic colonoscopy showed colonic polyps and colitis as the dependent variable, to combine five machine learning algorithms to build the prediction model (Fig. 11). We found the best prediction model is built by Nnet algorithm, and the AUC could achieve 0.706.

### 4.3.4 Galaxy Analysis Platform for Genomics Information

The genetic risk factors (GRF) are the pathogenic genes and regulatory factors of diseases and their major risk factors. They are of great significance in predicting the occurrence and development of diseases and would be the targets for gene therapy. The collection of the genomics information of patients is very important within our big data analytics platform. Then we have integrated the Galaxy analysis platform for supporting the import of genetic data. We can do a variety of bioinformatics analyses without downloading and installing any software or tools, and we can document every step of the analysis, while sharing the history of the analysis and the workflow of the construction with other

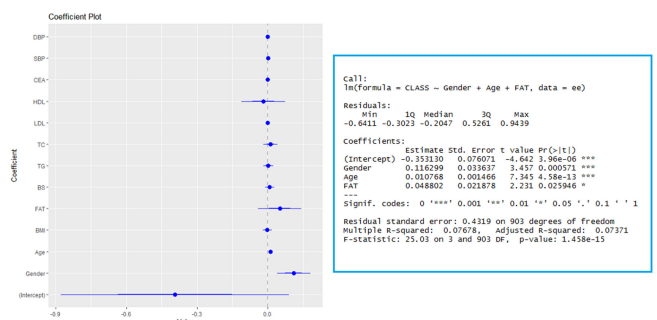


Fig. 10. Key factors related to the incidence of colonic polyps and inflammation, through stepwise regression equation and the correlation analysis.

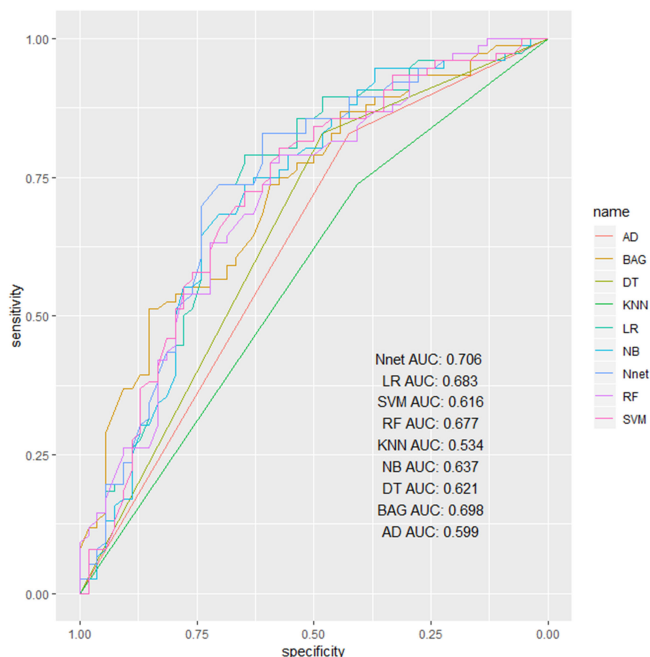


Fig. 11. The performance of colonic polyps and inflammation predict model and comparison of nine prediction model ROC curves.

researchers through the Galaxy analysis platform. Because of the construction and application of this platform are at the initial stage, the relevant genetic information of the patients is relatively little and still being collected and collated. We did not include the genomics information in this article. But this part of work is what we are doing and working towards.

## 5 DISCUSSIONS

Clinical data is the main source of healthcare big data, but with new technologies recently developed, it would be more valuable to have both clinical and verified general health, diets, sports, environments data integrated [18], [19]. Using the big data effectively, tangible research outcomes of digestive medicine could benefit many people's lives [20]. As digestive system diseases have great diversity and complexity, the digestive medicine analytic platform is suitable for complex analysis, the auxiliary of the digestive system disease early warning and prevention, early detection and treated, which has important clinical significance and value. It can also be beneficial to reduce healthcare costs in general and lower the burden of hospitals.

The conceptual model of digestive diseases proposed and the analysis case study shows the correspondence between the methodology and real-world practice. Based on the risk factors found from the research, paying enough attention to the fact of gastrointestinal continuum, carrying out the regular health check, risk assessment and individualized intervention enduringly, the occurrence of digestive system diseases could be lower or progress of diseases could be delayed with data-based guidance.

Study result recommends to carry out proper health examinations covering digestive system diseases as early as 40s and no later than 48. As most people have no obvious symptoms in their daily lives, the screening can help people

to know more about the status of their own bodies, plus the more people do health examinations, the more data could be collected for epidemiologic study, which could then support better in digestive medicine big data analysis for regional residents. In the case study presented, analysis was carried out in subjects who came to Xiangya Hospital for health examination with a colonoscopy from June 2011 to December 2015. Findings could then be beneficial to this group and similar groups of people for early detection and evaluation of gastrointestinal diseases.

It was found from the study that the positive rate of intestinal diseases was higher in the healthy people who had a health examination in Xiangya Hospital than other regions in the study. In terms of disease distribution, the morbidity of nonspecific colorectal inflammation was much higher than other regions. It might be related to the regional difference in the distribution of colorectal diseases, including people in Hunan province like to eat pepper and the economic development level [21]. Furthermore, data analysis shows that colonoscopy is of great value in the detection of digestive tract diseases. In addition, the incidence of digestive diseases was different between different age of groups. For example, the incidence of the precancerous lesions associated with gastrointestinal tumors and colorectal cancer increased with age. It is implied that early screening with colonoscopy is of great significance for the prevention and treatment of colorectal cancer. It is found that sex, age and obesity were contributed the most to the occurrence of colonic polyps and inflammation by machine learning and establishment of prediction model. The above study which analyzed the of health data of the physical examination population, provided a lot of meaningful research results and clinical information.

More stable and accurate clinical information can be provided for clinical reference by artificial intelligence computer quantitative data analysis model for the scientific research platform. This is different from the empirical qualitative analysis of clinicians. Furthermore, clinicians can combine the demographics, serology, image date, gene expression and other complex molecular biological information to make more accurate judgments about a patient's health, illness and possible prognosis, with the help of artificial intelligence technology powerful data analysis ability. It is helpful for the individualized clinical decision making. The application of the healthcare big data will bring revolutionary changes and unprecedented opportunities to medical research.

## 6 CONCLUSION AND FUTURE WORKS

A new and practical digestive medicine big data analytic platform is presented to support gastroenterology research in the Department of Gastroenterology of Xiangya Hospital Central South University China. This analytic platform adopted a distributed data center operating system to provide distributed scheduling and resource coordination functions, and enabled the data center resilient scalable software stack. It could schedule and manage all the data center resources as a single computer. The platform can help researchers to explore, analyze and mine data easily through self-service data analysis and mining using different machine learning algorithms. A real-world case study shows the effectiveness in analytic capability of the platform. Furthermore,

this platform has been gradually opened to other clinical departments to support different clinical research work in different diseases. And our plan is to build this big data platform into a multi-center clinical big data application platform, which will cover the most of the large and medium-sized hospitals in the whole Hunan province.

Future work would be done in the following aspects: 1) continuous work to enhance the standardized terminology knowledge base for different diseases with latest updates from clinical guides; 2) continuous work to extend the data collection work for digestive medicine big data for wider data coverage including health, weather, environment, diet, and sports of patients and the general public; 3) continuous work to improve the machine learning techniques and algorithms to offer better computational results for analysis.

## ACKNOWLEDGMENTS

The authors appreciate the comments and suggestions from anonymous reviewers that improved the clarity of this paper. This work is supported by the National Natural Science Foundation of China under Grant No. 81670589, the National Key R&D Program of China under Grant No. 2017YFC0909900 and Grant No. 2018YFC2002400, and the Key R&D Program of Hunan Province China under Grant No. 2017SK2013, and the Natural Science Foundation of Hunan Province under Grant No. 2019JJ40524.

## REFERENCES

- [1] M. Chen and N. Liu, "Research on the current status of development for healthcare big data," *J. Med. Inform.*, vol. 38, no. 7, pp. 2–6, 2017.
- [2] T. Dai, "Health and medical big data development perspective," *J. Med. Informat.*, vol. 37, no. 2, pp. 2–8, 2016.
- [3] N. h. a. f. p. c. s. i. c. website, "Action plan for promoting big data development Issued by the state council," *Chin. J. Health Inform. Manage.*, vol. 12, no. 5, 2015, Art. no. 447.
- [4] M. Chen and N. Liu, "Research on the development current state of healthcare big data," *Chin. Hospital Manage.*, vol. 37, no. 2, pp. 46–48, 2017.
- [5] M. Dai and Q. Meng, "Opportunities and challenges in data mining and data analysis of healthcare big data," *Chin. J. Health Inform. Manage.*, vol. 14, no. 2, pp. 126–130, 2017.
- [6] P. Wang *et al.*, "Needs analysis and platform building of medical big data application," *Chin. Hospital Manage.*, vol. 35, no. 6, pp. 40–42, 2015.
- [7] National health and family planning commission, "China health and family planning statistical yearbook (2013)," China Union Medical College Press, version 1, 2013.
- [8] F. Xue, "Healthcare big data-driven theory and methodology for health management," *J. Shandong Univ. (Health Sci.)*, vol. 55, no. 1, pp. 1–29, 2017.
- [9] R. Li and R. Lei, "Study on the value of digestive endoscopy in the diagnosis and treatment of early gastrointestinal cancer," *China Continuing Med. Educ.*, vol. 8, no. 3, pp. 100–101, 2016.
- [10] D. You and L. Cai, "The research of application prospect, challenges and countermeasures of the big healthcare data in Yunnan province," *Yunnan Sci. Technol. Manage.*, vol. 1, no. 1, pp. 14–16, 2017.
- [11] I. Masic *et al.*, "Evidence based medicine - new approaches and challenges," *Acta Informatica Medica*, vol. 16, no. 4, pp. 219–25, 2008.
- [12] S. Zhang *et al.*, "The current development situation and domestic countermeasures of evidence - based medical informatics," *J. Med. Inform.*, vol. 39, no. 10 no. 7, 0-54, 2018, Art. no. 5.
- [13] H. Ono *et al.*, "Guidelines for endoscopic submucosal dissection and endoscopic mucosal resection for early gastric cancer," *Digestive Endoscopy*, vol. 28, no. 1, pp. 3–15, Jan. 2016.
- [14] X. Wang, "Application design of clinical scientific research platform based on medical large data," *Inf. Med.*, vol. 17, no. 59, pp. 59–60, 2017.
- [15] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–54, 1996.
- [16] W. Chen *et al.*, "Cancer statistics in China," *CA: A Cancer J. Clinicians*, vol. 66, no. 2, pp. 115–132, 2016.
- [17] C. s. o. d. endoscopy and C. a.-c. a. Professional committee of oncoendoscopy, "Guidelines for early colorectal cancer screening and endoscopic diagnosis and treatment in China," *Gastroenterology*, vol. 20, no. 6, pp. 345–365, 2015.
- [18] R. Wyber *et al.*, "Big data in global health: Improving health in low - and middle - income countries," *Bull. World Health Org.*, vol. 99, no. 3, pp. 203–208, 2015.
- [19] Al. Pentland, T. G. Reid, and T. Heibeck, "Big data and health: Revolutionizing medicine and public health," WISH Big Data and Health Report, 2013.
- [20] X. Huang, X. Luo, P. Wang, and H. Wu, "Implementation progress of precision medicine-based on healthcare big data," *J. Med. Informat.*, vol. 38, no. 9, pp. 17–21, 2017.
- [21] Y. J. Surh and S. S. Lee, "Capsaicin, a double-edged sword: Toxicity, metabolism, and chemopreventive potential," *Life Sci.*, vol. 56, no. 22, pp. 1845–55, 1995.



**Lu Yan** received the MD degree from the Xiangya Hospital of Central South University, in 2013. She is now an attending physician with Xiangya the Hospital of Central South University. Her research interests include diseases of pancreas, diagnosis, and treatment of gastrointestinal tumors.



**Weihong Huang** received the BEng degree in automation and the MEng degree in pattern recognition and smart control from Southeast University China, in 1995 and 1998, respectively. He received the PhD degree in computer science from Nanjing University China, in 2001. From 2001 to 2002, he was a postdoctoral research fellow with the CNRS University Lyon 1 France. From 2002 to 2005, he was a lecturer with the Department of Computer Science, University of Hull, United Kingdom. From 2005 to 2014, he was a senior lecturer with the School of Computer and Information Systems, Kingston University London, United Kingdom. Since 2016, he has been appointed as a professor and the deputy director of the Mobile Health Ministry of Education - China Mobile Joint Laboratory, Xiangya Hospital Central South University, China. His research interests include mobile health, artificial intelligence in healthcare, cognitive computing for healthcare, semantic multimedia computing, and knowledge graph applications.



**Liming Wang** received the bachelor's degree from Liaoning Technical University, in 2010. He is now a data analyst with Bitvalue Technology (Hunan) Company Limited. His research interests include big data, data analysis, and artificial intelligence.



**Song Feng** received the bachelor's degree from East China Normal University, in 1994, and the master's degree in engineering from Central South University, in 2007. He is now a senior engineer with the Network Information Center of Xiangya Hospital, Central South University. His research interests include medical informatization, big data, data governance and analysis.



**Yonghong Peng** is a professor of data science and the leader for data science research with the University of Sunderland, United Kingdom. His research areas include data science, machine learning, data mining, and artificial intelligence. He is the chair for the Big Data Task Force (BDTF), and a member of Data Mining and Big Data Analytics Technical Committee of the IEEE Computational Intelligence Society (CIS). He is also a founding member of the Technical Committee on Big Data (TCBD) of IEEE Communications Society and an

advisory board member for IEEE Special Interest Group (SIG) on Big Data for Cyber Security and Privacy. He is an associate editor for the *IEEE Transaction on Big Data*, and an academic editor of *PeerJ* and *PeerJ Computer Science*.



**Jie Peng** received the MD and PhD degrees from the Xiangya Hospital of Central South University, in 2004. He is now a professor with the Xiangya Hospital of Central South University. His research interests are diseases of pancreas and the application of the health care big platform in the Gastroenterology Department.

▷ **For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/csdl](http://www.computer.org/csdl).**