Logistic big data prediction of gastric cancer

Yuxuan Wang¹, Caiwei Zhong², Yuxiang Zheng³

¹School of information engineering, Inner Mongolia University of Science and Technology, Baotou, Inner Mongolia, 014000
²Mathematical Science, University of Liverpool, Britain
³Medical College of Soochow University, Suzhou, Jiangsu, 215123

Abstract. In the era of big data, precision medicine for gastric cancer refers to the high throughput analysis of clinical data of patients with gastric cancer. The proposal of personalized diagnosis and treatment for each patient's molecular biology and pathological characteristics, so as to achieve personalized precision medicine. In this paper, the prediction method is significance analysis and the statistical method of prediction model is logical analysis and discriminant analysis, and the significance test standard is p<0.05. Logistic regression was used to identify and predict the prevalence of cancer patients, and the data used for prediction analysis were all from MS SQL database. The results showed that blood routine, blood biochemical and urine routine data can be used to distinguish cancer patients from healthy people. The cancer risk prediction model based on blood routine, blood biochemical and urine routine data can accurately target high-risk cancer people with an accuracy of 95.5%. In the era of big data, the accurate diagnosis and treatment strategy of gastric cancer can promote the aggregation, integration and sharing of gastric cancer related data.

1. Background

Gastric cancer is one of the most serious diseases threatening human life and health. The different characteristics of gastric cancer across countries and regions suggest that geographical diversity remains significant and that regional risk factors remain at different stages of social and economic transition. To reduce its morbidity and mortality, the national adjustment of public health resources, to carry out the positive and effective early detection treated early intervention, in 2004 and 2005, the ministry of public health and Chinese foundation and some provincial bureau jointly establish gastric cancer early detection treated early demonstration base, and implementation of promotion, early detection treated early gastric cancer as one of the main strategies of gastric cancer control, and puts forward specific measures, such as research and develop primary gastric cancer screening and early detection treated early guides, specification, high incidence in the rural areas respectively establish esophageal cancer, gastric cancer, liver cancer and early detection of nasopharyngeal carcinoma treated early demonstration base, etc.

1.1. Research object

Gastric cancer prevention and control measures of early detection treated early project involves many fields, large amount of information collection, the causes to explore such as prevention of early warning research questionnaire is various, and requires careful, at the same time, the technical level of the project sites is uneven, also cannot achieve the requirements of system integration, integrated information to
form a national, provincial, city and county (district) cooperative development of information network, need standardization and normalization. Regional health informatization is the trend in the development of medical industry, the meaning is very profound, it bodes well for the health care industry resources to make use of information technology to integrate a new direction, with the drive of information sharing will be changes in the medical service mode for people's health to provide more high quality, convenient, safe, efficient, and medical and health services. In gastric cancer early detection and prevention and control, information management system of treated early how to involve the scattered islands data collection, processing, integration, quickly form the stomach cancer prevention and control and make use of network technology resource information open new areas, there is a need for a digital revolution, the implementation of its resources informatization is the inevitable trend of future development of data, large data input in the development of information and the value of application is also very important.

In this study, blood routine, blood biochemistry and urine routine data were analyzed and predicted in 486,394 patients, including 40,217 patients with gastric cancer and 446,177 healthy subjects. Among the patients with gastric cancer, males accounted for 68% and females 32%. The average age of males was 61 years old and females 64 years old. The average age of men is 42 years old, and that of women is 40 years old.

1.2. Treatment of malignant tumors in the era of large data

At present, the treatment of malignant tumors is still dominated by traditional treatment methods. On the basis of standardized treatment for a large number of patients with malignant tumors, high-throughput analysis of the clinical data of patients with malignant tumors using big data can provide important information for the formulation of individual treatment plans. In addition, genomics and proteomics can be combined to analyze the gene and protein expression of patients, thus providing a large amount of effective data for precision medicine.

It is also suggested to establish a sustainable research cohort and corresponding biological sample bank to build a platform for promoting gastric cancer prevention and treatment research and resource sharing. To date more than a decade, the expert mode transform government model, technology model has been widely promotion, emerged a large number of excellent demonstration site, especially gastric cancer early detection in the rural areas treated early most estimation and project including 2018 upper gastrointestinal cancer, colorectal cancer, liver cancer, nasopharyngeal carcinoma and lung cancer, a total of 259 sites, covering the whole country except Taiwan, Hong Kong, Macao's 31 provinces, municipalities and autonomous regions, which existed in the digestive tract cancer in 194, 13 33 colorectal cancer, liver cancer, nasopharyngeal carcinoma and lung cancer, a total of 259 sites, covering the whole country except Taiwan, Hong Kong, Macao's 31 provinces, municipalities and autonomous regions, which existed in the digestive tract cancer in 194, 13 33 colorectal cancer, liver cancer, nasopharyngeal carcinoma, 8, 11, lung cancer of upper gastrointestinal tract cancer early detection treated early pattern of rural gradually promotion and application, Coverage of the population gradually increased and achieved good social benefits. In the early stage, it was implemented and promoted gradually based on the demonstration base, and the data collection process in various fields was established, which provided effective scientific data for the summary, analysis and research at that time. But with the increase of number of sites, an unprecedented expansion of scope of coverage, strengthen the implementation of the implementation of the technology and method of management of each link at the same time, the corresponding basic research data standardization, unity, timeliness, and expand the application in stand-alone mode have already can't meet the requirements of existing, forming islands of data will become the bottleneck of comprehensive utilization and development.

2. Forecast data

These new methods provide new ideas for the treatment of different stages of malignant tumors [2]. The study found that the key to early malignant tumor treatment lies in the individual factors of the patient. As a chronic disease, it usually takes several years for the mutated tumor cells to develop to the advanced stage of malignant tumor. The initial tumor cell development speed is slow, the body's immune cells can constantly find and kill tumor cells; With the increase of age, the decline of the body's immunity, as well as the influence of various factors such as the environment and itself, the cells undergo malignant
transformation, and the unskilled tumor cells are likely to metastasize to the whole body, resulting in the gradual formation of malignant tumors. However, due to individual differences, the time from the early stage to the late stage of malignant tumor is different, if the development is slow, it may take more than 10 years. If development is rapid, it may take only a few months or years to progress from early to late stage. The focus of early malignant tumor treatment is to prevent the further development of malignant tumor[3], and to alleviate the adverse reactions and pain of patients in the treatment process, so as to retain the physiological function of patients as much as possible, which is the strategy of precision medicine in the era of big data. Under the guidance of this thinking, need prior to treatment in patients with malignant tumor of the clinical characteristics and molecular biology, tumor biology behavior characteristics, comprehensive analysis to the patient's individual situation personalized treatment plan, and make an accurate judgment of the patient's disease process, thus further analysis of lymph node metastasis, blood line higher risk transfer and viscera transfer type, so as to expand the scope of lymph node cleaning, achieve the goal of cure of malignant tumor.

First has carried on the inspection to all the data analysis, SPSS analysis data has obvious time correlation, and considering the real meaning behind the data, abnormal, missing data can be derived from the original data measured data to predict, the time interval data is small, there is no fluctuation and irregular change, so the simple moving average method for time series prediction[2] and then realize the exception, missing data correction and repair.

In this study, the basic data covered in the routine health examination, including age, gender, height, weight, blood routine, blood biochemistry and urine routine, were adopted, a total of 48 items.

3. Logistic regression discriminant prediction

The prediction technique used in this study is Logistic Regression Analysis, which is widely used in medical research. At present, it is mainly used to screen risk factors in epidemiological studies[4], but it has good functions of discrimination and prediction at the same time. Especially, when the data types cannot meet the conditions of Fisher's discrimination and Bayes' discrimination, it shows the advantages and efficacy of Logistic regression discrimination.

\[
y = \log \frac{P}{1-P} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n
\]

Where \( y \) is the dependent variable, \( X \) is the independent variable, \( P \) is the probability, \( m \) is the intercept (constant), \( \beta \) is the regression coefficient, and Exp is the exponential function.

The risk assessment technique used in this study is net improved benefit equation (netliftalgmesh).

\[
Net \ lift = (Pt - Pc)
\]

Where, \( Pt \) is the percentage of cancer patients in the test group and \( Pc \) is the percentage of cancer patients in the control group.

The significance test criterion for statistical analysis and prediction in this study was \( p < 0.05 \) The statistical software used for statistical analysis and prediction is Python.

4. Independent verification of results

The prediction model of this study was built based on 4 years' data from 2010 to 2013. The 7 prediction models in the established prediction system will be independently verified by the data of 9,931 cancer patients and 110,077 healthy physical examinees from January to July 2014.

At present, there are few high-quality methods for early intervention in gastric cancer, and the cost-benefit ratio of advanced diagnosis and treatment is not high. The comprehensive prevention and
treatment effect of gastric cancer needs to be improved urgently, which not only brings great pain to patients, but also brings a heavy burden to the medical system. A lot of manpower, material resources and financial resources have been invested in the research of gastric cancer at home and abroad. Although progress has been made in clinical diagnosis \cite{5}, treatment and prevention, the etiology and pathogenesis of gastric cancer remain. There are many problems to be solved. Therefore, it is of great significance for the diagnosis and treatment of gastric cancer to fully grasp the regulation mechanism of gastric cancer. With the development of high-throughput technologies, researchers have been able to go deeper into the genome at an unprecedented depth and scale.

In the study. More than 90\% of the human genome can be transcribed to produce RNA, and there are less than 30,000 known protein-coding genes, less than 3\% of the transcription products. Most of the remaining transcriptional products are non-coding RNA (nc RNA) with unknown functions. Non-coding RNA is a kind of RNA that cannot be encoded into proteins, including transfer RNA (t RNA), ribosomal RNA Ribosomal RNA (r RNA), small interfering RNA (small interfering RNA, si RNA), small nucleolar RNA (small nucleolar RNA, sn RNA), and long non-coding RNA Long non-coding RNA (LNC RNA) and micro RNA (mi RNA), etc. The proportion of non-coding components in the genome of higher eukaryotes is significantly higher than that of lower prokaryotes, which is generally believed to be closely related to the complexity of gene expression regulation in higher organisms.

The expression of non-coding RNA is specific to cell types, tissues and developmental stages, especially in physiological and pathological processes.

![Fig 1. The formation of gastric cancer.](image-url)

At present, the studies on LNC RNA function mainly obtain observational data from the tissues or tumor cell lines of tumor patients, and make statistics on the differences in LNC RNA expression levels. Then, the research objects are assumed according to the degree of differences, and biological experiments are designed to verify the hypothesis. Established unfortunately, LNC RNA quantity is numerous, according to the NONCODE database, according to the established known human LNC RNA genes, there are about ninety-six thousand\cite{6}, because of the time-consuming traditional biology experiment, established the LNC RNA function studies tend to focus on a few established classical LNC RNA molecules, accounts for only about 1\% of the established known LNC RNA genes, established the vast majority of the remaining LNC RNA function is still not clear. With the rapid development of biotechnology and information technology, bioinformatics, as an independent discipline, has made great progress and has become an important discipline that no single discipline can replace, providing new opportunities for the prediction of tumor-related LNC RNA regulation.

As the incidence of gastric cancer in western countries is much lower than that in China and some Asian countries, western countries are much more concerned about the stomach. The study of cancer is relatively weak, and the northwest is the region with high risk of gastric cancer in China. Functional studies lack of effective exploration of the combination of bioinformatics and biological experiments. Therefore, the study of LNC RNA regulatory function prediction for gastric cancer based on bioinformatics method is important to provide accurate research direction and an effective and operable approach for the study of gastric cancer regulatory mechanism. Studies have shown that LNC RNA. The mechanism of competing endogenous RNA (ce RNA) plays an important regulatory role in the development of tumors. In the study of gastric cancer, the mechanism of LNC RNA in the regulation network of ce RNA in gastric cancer is still unclear due to the constraints of data sources, data processing methods, experimental conditions and other factors \cite{7}. In addition, although some researchers have found some LNC RNA that may be used as potential diagnostic and prognostic markers through
bioinformatics methods, LNC RNA is restricted by experimental technical conditions and other factors. Functional studies in gastric cancer often fail to integrate bioinformatics prediction and biological experiments effectively, leading to the lack of confirmation of the regulatory function of LNC RNA in gastric cancer. Therefore, gastric cancer LNC RNA. There are still many problems to be solved in the field of functional prediction.

5. Interpretation of result
This study on 2010-2013 of 30286 cancer patients and 336100 healthy check-up of 48 indicators related analysis and significance test, with significant, able to distinguish between cancer patients and healthy subjects established routine urine indexes of 7 kinds of single cancer risk prediction model (lung cancer, liver cancer, gastric cancer, colorectal cancer, esophageal cancer, breast cancer and cervical cancer, 7 kinds of cancer prediction model accuracy is over 95%\[8\], the average of 95.8%. The prediction model can predict 7 standard values between 1 and 100 for users. By comparing with the existing hematuria indexes of cancer patients, the prediction results can be dynamically analyzed to assess users’ cancer risk.

Due to the characteristics of various cancers, different cancers have different manifestations in conventional hematuria indexes, so different conventional hematuria indexes have different roles in predicting different cancers.

The occurrence and development of cancer\[4\] is a process from quantitative change to qualitative change. The changes of cancer cells are actually reflected in the hematuria index of human body. Since 90% of early cancers are asymptomatic, cancer patients do not show obvious symptoms or have no symptoms at all in the early stage, and only when the cancer cells have developed to a certain degree, the body will have a series of symptoms, so 80% of people once found cancer is already in the middle or late stage.

Fig 2. Schematic diagram of cancer growth treatment.
According to the results of this study, the cancer risk prediction system has three characteristics. Second, the prediction method is simple, based on the existing hematuria data, without further sampling. Third, forecasting costs are low, less than 10% of the market price.

6. Conclusion
Prevention is the final aim of we establish a cancer risk prediction system, the Irish professor paavo MERCYANDCORK university hospital has told us, "90-95% of cancer is caused by a lifestyle and smoking", our desire is: through early to predict the risk of cancer, the health and sub-health population issued a warning of "cancer", encouraging people to change their bad life style and quit, the last away from cancer.

By big data analysis of cancer risk assessment model can effectively use normal physical examination of blood routine, biochemical blood and urine, used for a variety of cancer risk prediction, and predict and verify the accuracy of more than 95%, which will provide cancer prevention and control of a kind of convenient, economic and effective new method, the early prevention of cancer treated early to play a positive role.

Reference
[1] Van Cutsem E, Sagaert X, Topal B, Haustermans K, Prenen H. Gastric cancer[J]. The Lancet, 2016, 388(10060): 2654-2664.
[2] Ajani J A, Lee J, Sano T, Janjigian Y Y, Fan D, Song S. Gastric adenocarcinoma[J]. Nature reviews. Disease primers, 2017, 3: 17036.
[3] Bray F, Ferlay J, Soerjomataram I, Siegel R L, Torre L A, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA Cancer J Clin, 2018, 68(6): 394-424.
[4] Chen W, Sun K, Zheng R, Zeng H, Zhang S, Xia C, Yang Z, Li H, Zou X, He J. Cancer incidence and mortality in China, 2014[J]. Chinese Journal of Cancer Research, 2018, 30(1): 1.
[5] Wan Y, Qu K, Zhang Q C, Flynn R A, Manor O, Ouyang Z, Zhang J, Spitale R C, Snyder M P, Segal E, Chang H Y. Landscape and variation of RNA secondary structure across the human transcriptome[J]. Nature, 2014, 505(7485): 706-9.
[6] Lander E S. Initial impact of the sequencing of the human genome[J]. Nature, 2011, 470(7333): 187-97.
[7] Mattick J S. RNA regulation: a new genetics?[J]. Nat Rev Genet, 2004, 5(4): 316-23.
[8] Flynn R A, Chang H Y. Long noncoding RNAs in cell-fate programming and reprogramming[J]. Cell Stem Cell, 2014, 14(6): 752-61.