Cluster Analysis of Cancer Incidence among Smoking Men in China

Mingyue Liao

The Affiliated High School of University of Electronic Science and Technology, Chengdu, China

2252962522@qq.com

Abstract. In recent years, the incidence rate and mortality of lung cancer have shown a rapid growth trend, and its pathogenic factors have also attracted widespread attention. At present, it is universally acknowledged that smoking has a strong correlation with lung cancer. Statistics show that 80% of lung cancer patients have smoking history. At the same time, it is also pointed out that quitting smoking in normal people can change the composition of intestinal flora; And in the normal population and patients with inflammatory bowel disease, the composition of intestinal flora between smokers and non-smokers is different. But now, in colorectal adenoma and colorectal cancer patients, the impact of smoking on intestinal flora is still unclear. To sum up, this topic plans to further study the relationship between smoking, intestinal flora and colorectal tumors. Cluster analysis was used to comprehensively evaluate and classify the incidence data of malignant tumors in 32 tumor registries in China, so as to further understand the regional distribution characteristics of malignant tumors in China. The results showed that compared with normal healthy people, the composition of intestinal flora in patients with colorectal cancer was different, and smoking could also change the intestinal flora. However, whether intestinal flora plays a role in promoting the occurrence and development of colorectal cancer by smoking needs further study.

Keywords: Smoking; Incidence of Tumor; Cluster Analysis.

1. Introduction

In recent years, the incidence rate and mortality of lung cancer have shown a rising trend in all countries of the world, especially in China, which has become the biggest cancer killer threatening people's health and life [1]. Lung cancer is one of the most common malignant tumors. A statistic shows that in the mid-1970s, the death rate of lung cancer in China ranked behind gastric cancer, esophageal cancer, liver cancer and cervical cancer, about 5.47/100000, which is at a relatively low level compared with other countries in the same period. The occurrence of sporadic colorectal cancer can be roughly divided into three pathways: adenoma adenocarcinoma pathway (including the serrated pathway caused by serrated adenoma), inflammation cancer pathway and denovo pathway [2]. Among them, colorectal adenoma is the main precancerous disease of colorectal cancer, accounting for 85%~90% of all precancerous diseases of colorectal cancer, or even higher. Colorectal adenoma can be divided into tubular adenoma, mixed adenoma, villous adenoma and serrated adenoma from the histological level, and their canceration rates are 0.61%, 5.02%, 14.96% and 5% respectively [3]. In the survey of lung cancer patients who went to Yunnan Cancer Hospital from 2005 to 2014, the pathological type, male female ratio, age, and lung cancer in high incidence areas of lung cancer were described. Among the pathological types, adenocarcinoma has occupied the first place in all lung cancer pathological types, followed by squamous cell carcinoma, small cell carcinoma, etc. The majority of male patients were squamous cell carcinoma, while the majority of female patients were adenocarcinoma, and the majority of cancer foci were distributed in the right lung, And the incidence of upper lobe is more than that of lower lobe [4].

Fundamentally speaking, the essence of complex diseases is the change of gene function, such as gene SNP and gene expression change. Gene SNP is the most common human heritable variation, accounting for more than 90% of all known polymorphisms [5]. However, not all individuals will change their gene function and develop into cancer. In this process, the interaction between genes and the environment is also involved [6]. There are more and more reports about digestive tract tumors. However, most researchers focus on finding new cancer-related genes and studying the relationship

between these new genes and cancer occurrence and development, but ignore the interaction between genes and environment [7]. The purpose of this study is to use the statistical method of clustering analysis to comprehensively evaluate and classify the registry, verify and find the high incidence areas of malignant tumors, so as to understand the current situation and distribution of malignant tumors in China, provide important clues for further analysis and epidemiological research, and provide reference and decision-making basis for local development of cancer prevention and control strategies. The purpose is to describe the harm and risk of malignant tumors caused by smoking, and provide evidence-based data support for further reducing the smoking rate, creating a smoke-free environment and reducing the burden of tobacco related diseases.

2. Research Contents and Methods

2.1 Research Object

The Global Burden of Disease Study in 2015 shows that cancer is the second leading cause of human disease death in the world. Between 2005 and 2015, some cancer patients died of digestive tract cancer [8]. Among them, the total number of deaths of colorectal cancer reached 832,000, gastric cancer 818.9,000, liver cancer 810.5,000 and esophageal cancer 439,000. Most tobacco carcinogens first form DNA adducts through metabolic activation; DNA adducts accumulated in the body can induce some key genes to mutate, thus initiating cancer. The removal of DNA adducts requires DNA repair, which indicates that the variation of DNA repair genes may be related to the repair efficiency of different DNA damages [9]. In addition, there is competition among different detoxification pathways, and different individuals have different balance between metabolic activation and detoxification, thus affecting the risk of cancer [10]. For the tumor cases diagnosed for the first time by various medical institutions at all levels in the city, the outpatient or resident doctors shall fill in the Tumor Report Card, and the medical technology departments (pathology, CT/MRI, radiology, B-ultrasound/color ultrasound, endoscope and blood testing center) shall establish the Tumor Register to timely record the information related to the diagnosis and treatment of malignant tumor cases in their departments. The information of the Tumor Register was extracted from the medical technology departments and submitted to the District CDC for introduction into the unified tumor registration information management system of the city; If there is no hospital information system, the medical staff shall fill in the Tumor Report Card and timely enter it into the unified tumor registration information management system of the city. The 5-year data processing at the same monitoring point was used to analyze the fluctuation of the 5-year incidence rate. The National Cancer Registration Center used the database and analysis software to calculate the proportion of pathological or histological diagnosis, the proportion of only death medical certificates, the death/incidence ratio, the proportion of other or unspecified sites and unknown primary sites (secondary) malignant tumors, the incidence (death) rate Standardized the morbidity (13 deaths) rate, constituent ratio and other indicators, comprehensively reviewed and evaluated the integrity and effectiveness of the registration data, and finally accepted the data from 32 registries with good quality according to the audit results.

2.2 Experimental Methods

For those who meet the above enrollment criteria and have no exclusion criteria, the stool samples of patients who are going to undergo colonoscopy will be collected before their colonoscopy bowel preparation, and those who are going to undergo surgery will be collected before their bowel preparation and at least 2 weeks before the last colonoscopy (to exclude the influence of colonoscopy on intestinal flora). Before collecting the specimen, the subject was instructed to empty the bladder to avoid other body fluids polluting the experimental specimen. After collection, the specimen was placed in a sterile vessel, marked on the bottle cap and tube wall of the vessel, and immediately stored in the negative 80℃ environment until the experimental operation. In order to obtain the most suitable DNA concentration and purity, the following experimental steps have been fine-tuned according to the kit instructions and the results of many pre-experimental operations. The purity and concentration

of the extracted DNA all meet the 16Sr DNA sequencing conditions. (1) prepare reagents according to the requirements of QIAamp DNA stool Mini Kit. (2) Weigh 180-220mg feces, put it into a 2ml sterilized centrifuge tube, add 1.4ml Buffer ASL, and shake until the specimen is homogenized. (3) Centrifuge again for 3 minutes (14000rpm, 25℃), then suck 0.4ml of supernatant into a new 1.5ml sterilized centrifuge tube, and discard the residual liquid. (4) Shake the extracted DNA well, then use spectrophotometer to determine the DNA concentration and purity, and agarose gel electrophoresis (agarose gel concentration 1%, voltage 5V/min, time 20min) to detect the DNA integrity. Pan/Core species analysis is mainly used to describe the changes of total sample species and core species with the increase of sample size. It is used to analyze and judge that the sample size is reasonable and sufficient at the level of domain, kingdom, phylum, class, order, family, genus, species and OTU. When the curve tends to be flat, it indicates that with the increase of sample number, the total species number does not increase significantly, and the core species number does not decrease significantly, that is to say, the sample number is sufficient and reasonable. The incidence rates of 12 major cancers are classified by hierarchical clustering method, and the distance between classes is determined by Ward's method. In order to eliminate the influence of large fluctuations of variation coefficients of different variables on the clustering results, it is necessary to standardize the data in cluster analysis, and Z scores is used here for standardization. The specific clustering process is completed by SPSS13.0.

AIC (m) is a discrete function with respect to order m. The optimal order should minimize the value of the function. The method recommended by Schwart (1987) is adopted to select the order m. The maximum value of m is 41 (10012)/T (T represents the sample volume). The maximum value of m is 9. By establishing a linear regression model between the time series of lung cancer mortality and aerosol optical extinction coefficient, the AIC value calculation formula is adopted, and then the AIC values of 1-9 are calculated respectively. The results are shown in Figure 1.

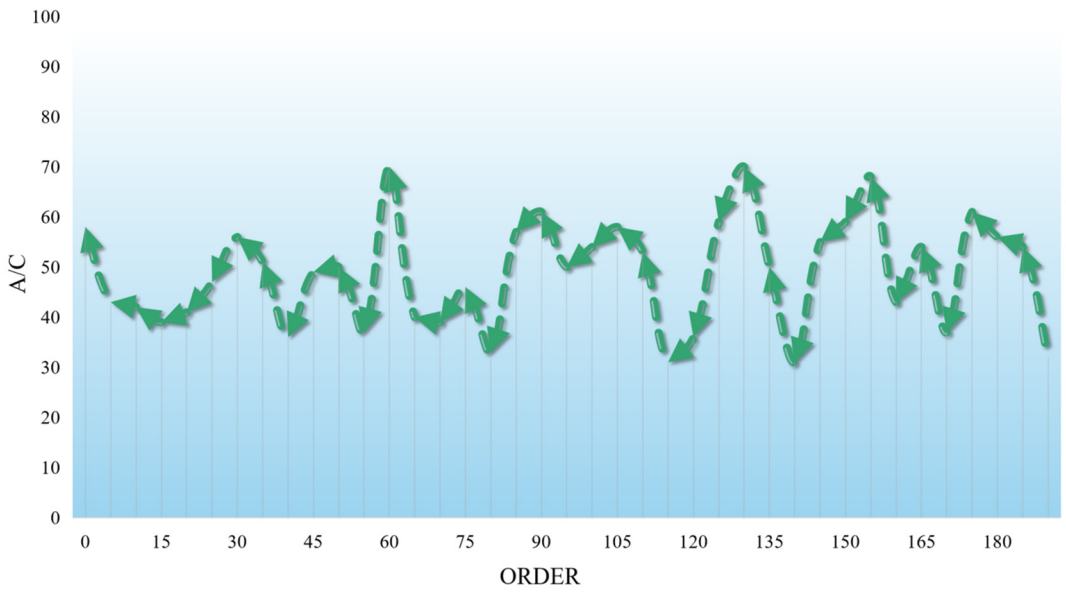


Figure 1. Relationship between AIC value and order

According to Figure 1, when the model order is 9, the AIC value reaches the minimum. Therefore, the best order of the linear regression model for the time series of lung cancer mortality and aerosol optical extinction coefficient is 9.

3. Causality Analysis of Smoking-induced Lung Cancer

3.1 Comprehensive Cluster Analysis Results of Main Cancer Incidence

Among the top five malignant tumors, the incidence of smokers is higher than that of non-smokers, and the difference is statistically significant. Especially in new cases of lung cancer, the incidence risk of smokers is 3.26 times that of non-smokers (OR = 3.26,95% CI: 3.14 ~ 3.38, P<0.01). Not only lung cancer, but also malignant tumors of digestive tract and urinary system. The risk of smoking is significantly higher than that of non-smokers, which indicates that smoking is closely related to malignant tumors of multiple parts and systems. When smoking, most of the smoke reaches the lungs, and a small part enters the digestive tract with swallowing, acting on the digestive system. There are about 69 known carcinogens in cigarette smoke, among which nicotine, carbon monoxide and tar are the most harmful. Besides smoking leading to lung cancer, which has been confirmed by a large number of epidemiological studies, it is also closely related to the incidence of liver cancer, stomach cancer, colorectal cancer and bladder cancer. According to the stratified analysis of smoking status, we found that there was no statistical difference in the influence of CYP1A1IIe462Val polymorphism on the susceptibility of four digestive tract cancers among non-smokers (OR=0.973,95%CI0.827-1.145; I2=0, heterogeneity test p=0.478). There was no statistical difference in heterogeneity between the studies. In the smoking population, there was no statistical difference between the two associations (OR=1.341,95% CI 0.959-1.876; I2=59.853%, heterogeneity test p=0.015), and there was moderate heterogeneity between this study. Through sensitivity analysis (Figure 16B), we found that when Fernandes, 2016 [195] study was excluded, the combined effect value OR of the remaining studies was 1.479 (95% CI: 1.181-1.853, I2=18.382%, heterogeneity test p=0.290), which was statistically significant. There was also no statistical difference between nonsmokers (OR=0.953,95% CI 0.694-1.310; I2=31.851%, p=0.185 for heterogeneity test) and smokers (OR=1.077,95% CI 0.75-1.547; I2=53.067%, p=0.047 for heterogeneity test), but there was heterogeneity between the two groups. The most obvious is that incidence rate Province has the highest incidence of nasopharyngeal carcinoma in China, and its incidence rate level is far higher than that of other parts of the country. Sihui City, Zhongshan City and Guangzhou City are grouped together. The distribution of liver cancer in China is higher in the southeast than in the northwest, and higher in coastal areas than in inland areas.

3.2 Analysis Results

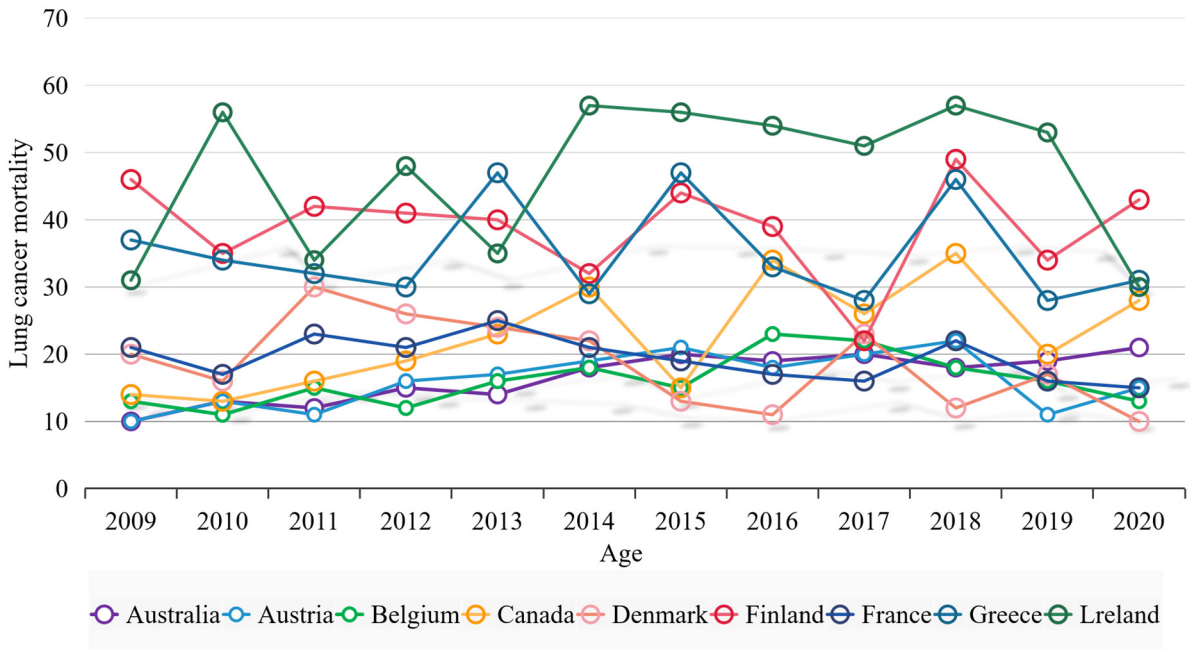


Figure 2. Smoking in each country

Although almost all studies have shown that smoking significantly affects the incidence rate of lung cancer, no study has shown the causal relationship between the two variables, that is, we cannot say that smoking causes lung cancer accurately. Due to the lack of scientific and systematic causal criteria, people pay more attention to the macro evaluation of the impact of smoking on lung cancer. Therefore, we cannot comprehensively and scientifically confirm the causal relationship between the two, and the results can only be probability or some specific relationship. According to the statistics of the Global Cancer Observatory, the standardized data of male lung cancer mortality (1/100000 people) and female lung cancer mortality in various countries were obtained. Since the change of smoking rate is based on the change trend of the average number of cigarettes smoked per adult per day, the sum of the lung cancer mortality of men and women is divided by 2 as the standardized average lung cancer mortality. The results are shown in Figure 2. With the increase of time, the lung cancer mortality first shows an increasing trend and then a decreasing trend.

The statistics of smoking amount are from Barbara Forey, Jan Hamling, Peter Lee and others in the international smoking statistics. Here we take the average daily smoking amount of each adult as the measurement of smoking degree. Since the two time series for calculating causality must be continuous, countries with discontinuous lung cancer mortality and smoking data and countries with lung cancer mortality but no smoking rate statistics were removed, and finally smoking data of 12 countries including Australia, Austria, Canada, the United States, and the United Kingdom were obtained.

To sum up, there are significant differences between smoking group and non-smoking group, mainly at the level of Magnesia and Fusobacterium. Among them, there was no significant difference between HC group and CRC group, while Fusobacterium was significantly increased in CRC group. Previous studies have also shown that the content of Fusobacterium in colorectal cancer tissues is significantly higher than that in adjacent tissues, including Fusobacterium nucleatum, Fusobacterium necrophorum and Fusobacterium necrophorum. As a typical representative of Fusobacterium, a large number of literatures have reported that Fusobacterium nucleatum is closely related to the occurrence and development of colorectal cancer. It has been reported that the level of Fusobacterium nucleatum in colorectal cancer tissues, feces and intestinal mucosa swabs is significantly higher than that in the normal control group, especially Fusobacterium nucleatum. The mortality series of smoking and lung cancer in different lag years, and estimate the corresponding linear regression model, then calculate Granger causality value and new causality value. For each method of causality, the number of years with the largest lag of causality is considered as the number of years with the lag of smoking on the mortality of lung cancer in this country. The results of the new causality show that the lag years of smoking on lung cancer mortality in 12 countries obey the normal distribution with an average of 44. However, the result of Granger causality is obviously incorrect, because the causal relationship between smoking and lung cancer in a country lag only 2 years. Above, we proved the relationship between smoking and lung cancer by the method of causality. Smoking not only does harm to the health of individuals, but also the victims of second-hand smoke abound, and some studies have shown the harm of passive smoking. As described by the American Centers for Disease Control and Prevention, smoking is the most important preventable factor of human health. We should take appropriate measures to reduce the smoking rate, such as increasing the tax rate of tobacco and implementing the tobacco control order.

4. Conclusion

The incidence rate and mortality of lung cancer have shown a rapid growth trend worldwide. In recent years, lung cancer has become the first killer of human health. Male smokers have obvious differences in the epidemic characteristics of tumor incidence between urban and rural areas, that is, the urban incidence rate is higher than that in rural areas, while the age of incidence in rural areas is earlier than that in urban areas. Therefore, the education and intervention of tobacco control should not neglect and weaken rural areas. We should increase the efforts of smoking cessation education in

rural areas, actively explore ways of tobacco control suitable for rural areas, and improve the smoking cessation rate of the population and the coverage rate of media information on tobacco control. The incidence of cancer in the world is increasing by 3% every year, and the incidence and mortality of cancer in China are also on the rise. In view of the great threat posed by cancer to human health and social sustainable development and the importance of cancer registration, it is necessary to further improve the national cancer registration system, increase the number of registration points, and expand the monitoring scope. In particular, we should pay attention to the resource investment in the registration work in the western and rural areas with high incidence of cancer, do a good job in early diagnosis and treatment of cancer, and strive to reduce the incidence and mortality of cancer.

References

[1] Sun H X, Zhang M X, Wang W Y. Analysis of cancer incidence and mortality in Heilongjiang cancer registries, 2016[J]. TMRTumors, 2020, 3(4):8-10.

[2] Xu L, Ma Y, Wang S. Incidence of gastrointestinal stromal tumor in Chinese urban population: A national population-based study[J]. Cancer Medicine, 2020, 10(2):12-15.

[3] Daniel J A, Jin K M, Poudel K C. Family Support and Readiness to Consider Smoking Cessation among Chinese and Vietnamese American Male Smokers. [J]. The Journal of Smoking Cessation, 2021,26 (10): 15-18.

[4] Shakeri M T, Nezami H, Nakhaee S. Assessing Heavy Metal Burden Among Cigarette Smokers and Non-smoking Individuals in Iran: Cluster Analysis and Principal Component Analysis[J]. Biological Trace Element Research, 2021(3):1-9.

[5] Smit E S, Brinkhues S, Vries H D. Subgroups Among Smokers in Preparation: A Cluster Analysis Using the I-Change Model[J]. Informa Healthcare, 2018(3):20-22.

[6] Konno S, Taniguchi N, Makita H. Distinct Phenotypes of Smokers with Fixed Airflow Limitation Identified by Cluster Analysis of Severe Asthma. [J]. American Thoracic Society, 2018,23(1):26-30.

[7] Konno S, Taniguchi N, Makita H. Distinct Phenotypes of Smokers with Fixed Airflow Limitation Identi?ed by Cluster Analysis of Severe Asthma[J]. Annals of the American Thoracic Society, 2017: AnnalsATS. 2017,62(2):12-14.

[8] Zhang C, Jiang C X, Zhong Q. The status and trends in malignant tumor incidence rates from 2011 to 2014 in areas along the Yangtze River of Anhui Province[J]. Chinese Journal of Disease Control & Prevention, 2018,33(7):25-30.

[9] Ding X B, Xiao-Yan L V, Mao D Q. Quantitative analysis of incidence trend of malignant tumor and its influencing factors in Chongqing Municipality[J]. South China Journal of Preventive Medicine, 2017, 66 (8): 28-32.

[10] Handebo S, Birara S, Kassie A. Smoking Intensity and Associated Factors among Male Smokers in Ethiopia: Further Analysis of 2016 Ethiopian Demographic and Health Survey[J]. BioMed Research International, 2020, 2020(1):2-8.