Bayesian Estimation of the Risk of Covid-19 Infection in Data Constraints

Rui Wu*, Lisong He Yan Wang and Qian Cheng

School of Finance, Xi'an Eurasia University, Xi'an, China

*Corresponding author e-mail: wurui@eurasia.edu

Keywords: Covid-19, Bayesian estimation, Chinese center for disease control and prevention

Abstract: Based on the epidemiological data and the related population data published in the research paper of the epidemiological group of the Chinese Center for Disease Control and Prevention(CDC) on the mechanism of emergency response to coronavirus pneumonia, the risk of coronavirus infection in returning students is estimated by using the Bayesian principle. The results show that the Bayesian principle can be used to estimate the risk of students' infection according to the prior probability under the data limitation.

1. Introduction

This paper explores the problem of risk identification of coronavirus infection under data limitations. The coronavirus broke out in late 2019, and by late May more than five million people worldwide had been infected and more than 300, 000 had died. China attaches great importance to quarantine work, deploys closely and controls over the epidemic, and has achieved remarkable results. However, with the re-study of various places, the risk of epidemic situation can not be ignored, especially the re-study of universities everywhere, the area of university students is large, after returning to school concentrated accommodation, how to estimate the risk of epidemic emotion infection of returning to school students, so as to manage the students with different risk degree in advance, so as to realize the orderly re-study and avoid the rebound of epidemic situation after returning to school.

Since the outbreak of the epidemic, scholars have highlighted the important role of big data to help the epidemic prevention. For example, Wu He Quan (2020)put forward the idea of mobile phone, electronic equipment, data monitoring, medical prevention and control under the big data, Liu Guanghao (2020) introduced China's basic telecommunications enterprises, Internet enterprises and related research institutions using technology and data advantages, in the epidemic situation awareness, migration path tracking, re-production, etc But this part of the research, more summary analysis, less model analysis, especially many universities do not have large data detection system. At present, the scanning green code inspection used everywhere can be regarded as an application of the large data detection system, mainly using the user's residence and other information, but the information considered is not comprehensive.

Logically, if a part of the database of infected patients can be obtained, such as the large-data-assisted coronavirus prevention scenario, with the specific information of infected people, cured people, uninfected people, then the risk estimation of coronavirus infection can be solved by supervised classifier algorithm, such as logical regression, decision tree, etc. But because of information limitations, each research department may have only partial data, making risk assessment more difficult. Especially in colleges and universities, it is difficult to obtain these data in the prevention and control of epidemic situation. This paper discusses the problem of infection risk identification under data limitation. Based on the epidemiological characteristics data published in the research paper of the epidemiology group of the new coronavirus pneumonia emergency response mechanism of the Chinese Center for Disease Control and Prevention, combined with the related population, occupation and age data published by the National Bureau of Statistics, the Bayesian principle is used to estimate the risk of infection and provide some suggestions for the prevention and control of the epidemic situation in colleges and universities.

  DOI: 10.25236/cseem.2020.033

2. Introduction to Bayesian Principle

The Bayesian principle was proposed by Bayes in the 18th century.

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

If A is an event, then P(A) is an estimate of the likelihood of event occurrence in advance, and P(A|B) is the probability of event occurrence in the event B. When the conditional events are multiple, the Bayesian formula is as follows:

$$P(A \mid B_1, B_2, \dots, B_i) = \frac{P(B_1, B_2, \dots, B_i|A)P(A)}{P(B_1, B_2, \dots, B_i)} \quad (2)$$

$P(B_1, B_2, \dots, B_i)$and $P(B_1, B_2, \dots, B_i|A)$ are the joint probability of multiple events, it is often difficult to estimate, and it is generally assumed that the conditions are independent between events.

$$P(A \mid B_1, B_2, \dots, B_i) = \frac{P(A)\prod P(B_i|A)}{\prod P(B_i)} \quad (3)$$

Because the assumptions are strong, they are also called Naïve Bayes. In some cases, the hypothesis of conditional event independence is difficult to satisfy, for example, when used to estimate the epidemic risk, some concurrent characteristics of patients are not independent, but these characteristics of non-infected persons are independent, so Bayesian formula can be transformed into:

$$P(A|B_1, B_2, \dots, B_i) = 1 - \prod(1 - P(A|B_i) \quad (4)$$

For example, the mark Y=0 is not diagnosed, Y=1 is diagnosed, and a plurality of characteristics are estimated by the formula 4. For example, the calculation formula is:

$$P(Y = 1 \mid \text{History of exposure in Wuhan, male})$$
$$= 1 - P(Y = 0 \mid \text{History of exposure in Wuhan, male})$$
$$= 1 - P(Y = 0 \mid \text{History of exposure in Wuhan,}) P(Y = 0 \mid \text{History of exposure in Wuhan})$$
$$= 1 - (1 - p(Y = 1| \text{ History of exposure in Wuhan })(1 - p(Y = 1|male)) \quad (5)$$

3. Risk Estimation of Virus Infection Based on Bayesian Principle

According to Bayes theorem, we can infer the infection probability of a new sample according to the probability of different cases, even if there is no specific disease data of each individual. The specific methods of estimation are as follows:

(1) Firstly, according to the epidemiological data published in the research paper of the epidemiology group of the Chinese Center for Disease Control and Prevention(CDC)on the mechanism of emergency response to coronavirus pneumonia, the characteristics of the samples are determined: whether there is a history of exposure in Wuhan, whether people are in contact with Wuhan recently, and the probability of infection under different characteristics. As shown in Table 1.

Table 1 Epidemic Characteristics Data Released by the Epidemiology Unit of the Emergency Response Mechanism of the New Type of Coronavirus Pneumonia

Characteristics	detailed characteristic	confirmed cases	Proportion (%)	probability
age group	0-	416	0.9	0.009
	10-	549	1.2	0.012
	20-	3619	8.1	0.081
	30-	7600	17	0.17
	40-	8571	19.2	0.192
	50-	10008	22.4	0.224
	60-	8583	19.2	0.192
	70-	3918	8.8	0.088
	>=80	1408	3.2	0.032
sex	male	22981	51.4	0.514
	female	21691	48.6	0.486
Occupation	Services	3449	7.7	0.077
	Farmers/workers	9811	22	0.22

	Medical personnel	1716	3.8	0.038
	Retired persons	9193	20.6	0.206
	Other	20503	45.9	0.459
Provinces	Hubei	33367	74.7	0.747
	Other	11305	25.3	0.253
Wuhan Exposure History	Yes	31974	85.8	0.858
	No	5295	14.2	0.142
	Loss	7403		
Basic diseases Hypertension	Hypertension	2683	12.8	0.128
	Diabetes	1102	5.3	0.053
	Cardiovascular diseases	873	4.2	0.042
	respiratory infectious disease	511	2.4	0.024
	Cancer	107	0.5	0.005
	No	15536	74	0.74

(2)Collect and collate the basic data of the population of the whole country, the provinces, the cities, the industries and the occupations corresponding to the above-mentioned related indexes, and calculating the prior probability, as shown in Table 2.

Table 2 Basic Population Data

Large category	Category	Population	Data sources	Number of confirmed cases	conditional probability formula	conditional probability
National	Total population	1400050000	National Statistical Office	44672	Number of confirmed cases/total population	3.19074E-05
	Male population (10, 000)	715270000	National Statistical Office	22981	Male confirmed/total male population	3.21291E-05
	National female population (10, 000)	684780000	National Statistical Office	21691	Women confirmed/total female population	3.16759E-05
Hubei	Population in Hubei	59270000	National Statistical Office	33367	Number of confirmed cases in Hubei /total population in Hubei	0.000562966
	Population outside Hubei	134078000 0	National-Hubei	11305	Number of confirmed cases outside Hubei /total population outside Hubei	8.43166E-06
Industry	Employment in the First Industry	19445.2	National Statistical Office			
	Employment in the secondary sector	21304.5	National Statistical Office			
	Employment in the tertiary sector	367213	National Statistical Office			
	Retired persons	175990000	Population aged 65 and over	9193	Number of confirmed retirees/retirees	5.22359E-05
	Other	436422000	Total population minus four categories	20503	Other confirmed/other population	4.69798E-05

(3)Questionnaire survey will be conducted for returning students to collect data on each person's relevant characteristics.

Table 3 Student-Related Characteristics Questionnaire

serial number	Issues
1	Your sex
2	Your current location
3	Has the holiday been to Hubei
4	Have you ever been to Hubei before contact with family or other people

5	Parental occupation (probability of alternative occupation)
6	Number of family members returning workers
7	Whether fever, whether chest tightness, whether respiratory discomfort

(4)Based on the student survey data, combined with the probability calculated in Table 2, the probability of each student's virus infection is inferred based on the Bayesian principle.

4. Empirical Study on Risk Estimation of Virus Infection Based on Bayesian Principle

This study collected the data of 179 students in this specialty by means of questionnaire. The results were listed in Table 4 below.

Table 4 Survey Results Of Students on Return to School

Large category	Category	Percentage (%)
Sex	Male	32
	Female	68
Place of family	Shaanxi	69
	Other	31
Physical condition	Is there heat	0
	Are you sulky	0
	Is respiratory discomfort	0
	No more than symptoms	100
Parental career	Services	6
	Peasant workers	43
	Medical personnel	4
	Retired persons	4
	Other	43
Have you been to Hubei	Yes	98
	Yes	2
Is there any exposure history in Wuhan	Yes	98
	Yes	2

In terms of sex, men are more likely to be in good health and have no symptoms such as fever. Because the school is located in Xi'an, so families are located in Shaanxi more. In terms of occupations, 43 per cent of parents are farmers and workers, 43 per cent are others, 4 per cent are health workers and retirees and 6 per cent are services. The percentage of people who went to Hubei was 25, and the percentage of people who had been exposed to Wuhan was 2%.

According to the formula 5, the risk of each student infected with coronavirus is estimated according to the survey data and the prior probability, and after the estimated probability is sorted, the statistics of the student estimated risk are shown in Table 5.

Table 5 Statistical Table For Description of Risk Estimates

minimum	median	mean	maximum
4.488e-05	8.247e-05	1.750e-04	2.940e-03

On the whole, the probability of these students infected with coronavirus was not high, and the maximum risk of infection was only 2. 940 e-03.

5. Conclusion

Based on the Bayesian principle, the epidemiological data published in the research paper of the epidemiological group on the mechanism of emergency response to coronavirus pneumonia in the Chinese Center for Disease Control and Prevention are used to collate the data and calculate the prior probability. The risk of new coronavirus infection was estimated after a questionnaire was conducted among returning students. The results show that the Bayesian principle can be used to estimate the risk of students' infection according to the prior probability under the condition of lack of specific data, and can solve the problem of risk identification for returning students to university to some extent.

References

[1] LIU Guanghao. Inspiration from the experience of big data at home and abroad in preventing and controlling the epidemic[J]. China Telecom, 2020(04):68-71.

[2] Ou Hequan. Big data to help prevent and control epidemic situation[J]. Big data era, 2020(03):26-33.

[3] Kupferschmidt, K. (2020). Study claiming new coronavirus can be transmitted by people without symptoms was flawed. Science | AAAS. doi: 10.1126/ science.abb1524.

[4] Epidemiological Group of Emergency Response Mechanism of New Type of Coronavirus Pneumonia in Chinese Center for Disease Control and Prevention. Epidemiological Characteristics of New Type of Coronavirus Pneumonia[J/OL]. Chinese Journal of Epidemiology, 2020, 41(2020-02-17).