**An evaluation model of 2019-nCoV prevention and control effectiveness based on immigration data in China**

Qiangsheng Huang

**Abstract**

**Background:** As of the end of February 2020, 2019-nCoV is currently well controlled in China. However, the virus is now spreading globally.

**Objective:** This study aimed to evaluate the effectiveness of outbreak prevention and control measures in a region.

**Methods:** A model is built for find the best fit for two sets of data (the number of daily new diagnosed, and the risk value of incoming immigration population). The parameters (offset and time window) in the model can be used as the evaluation of effectiveness of outbreak prevention and control.

**Results:** Through study, it is found that the parameter offset and time window in the model can accurately reflect the prevention effectiveness. Some related data and public news confirm this result. And this method has advantages over the method using R0 in two aspects.

**Conclusions:** If the epidemic situation is well controlled, the virus is not terrible. Now the daily new diagnosed patients in most regions of China is quickly reduced to zero or close to zero. Chinese can do a good job in the face of huge epidemic pressure. Therefore, if other countries can do well in prevention and control, the epidemic in those places can also pass quickly.

Key words: novel coronavirus (2019-nCoV)–infected pneumonia (NCIP); Epidemic Prevention and Control; Big Data Analysis; Data mining; Mathematical modeling

**Introduction**

It has been 2 months since the outbreak of novel coronavirus (2019-nCoV)–infected pneumonia (NCIP) in Wuhan, Hubei, since December 2019. Chinese has been ignorant of it from the beginning. For example, the people of Wuhan held a banquet for tens of thousands of people in January 18th, which caused the spread of the epidemic [1]. Then, Chinese gradually paid attention to the investigation and control of incoming immigration population, and closed down some densely populated regions. And minimize internal people going out, wearing masks when going out, community prevention, and so on [2,3]. At present, the epidemic in the regions other than Hubei, which is the region where the virus originated, have been basically controlled.

This study uses immigration population data and daily new diagnosed data to find ways to assess the evaluation for effectiveness of epidemic prevention and control in each region.

**Methods**

**Overview**

Because the spread of the virus is closely related to immigration population, the evaluation for effectiveness of prevention and control is not simply a matter of the number of daily new diagnosed. It should also consider on the pressure of the number of incoming potential virus spreaders in the past.

Therefore, this study uses two sets of raw data: “daily new diagnosed in each region” and “incoming immigration population”. And the third sets of data is constructed from the first two sets of data, that is the risk of daily incoming immigration population in each region.

A model is built in this study, which uses “offset” and “window” parameters for the best fit for “daily new diagnosed” and “daily immigration risk”. And these two parameters serve as an evaluation for the effectiveness of external and internal epidemic prevention and control. The model workflow is shown in figure 1. This section performs the corresponding theoretical analysis and model parameter description step by step.

**Abbreviations**

To simplify the description, the abbreviations are used for the following nouns in the later paper:

“The provinces / municipalities of China” is referred to as “region”.

“The daily new diagnosed in each region” is referred to as “daily new diagnosed” or just “new” (in quotes).

“The daily incoming immigration population size in each region” is referred to as “daily incoming population size” or just “size” (in quotes).

“The risk of daily incoming immigration population in each region” is referred to as “daily immigration risk” or just “risk” (in quotes).

“The offset parameter in the model” is referred to as “offset” (in quotes).

“The time window parameter in the model” is referred to as “window” (in quotes).

**Data used**

***Data 1: daily new diagnosed in each region***

Since January 2020, various provinces / municipalities in China have successively announced data such as the number of new confirmed diagnoses of the epidemic each day. The “daily new diagnosed” data is crawled from Doctor Lilac Network (<https://ncov.dxy.cn/>).

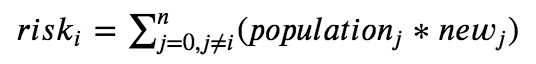
***Data 2: daily incoming immigration population size in each region***

The epidemic situation is exactly during the China Spring Festival transportation and its return journey. The population migration factor and its control must be the most important factor affecting the developing trend of the epidemic situation. Therefore, the daily incoming immigration population data, which are distinguished from different source regions, is crawled from Baidu Migration (<https://qianxi.baidu.com/>).

Because Hong Kong, Macao, and Taiwan due to unavailability of immigration population data, and Hubei because of the origin of the virus and its traffic closure leads to little incoming immigration population from late January, these regions haven’t been analyzed. However, the immigration population moved from Hubei to other regions where are analyzed are considered.

***Data 3: daily immigration risk in each region***

It should be considered that the immigration population with same size from different sources, its degree of risk is different. For example, early in the spread of the virus, people from Hubei, where is the birthplace of the virus, have a greater risk than from other regions with the same immigration size. Therefore, process the incoming immigration population as follow, and get the “daily immigration risk” data.



In the formula, all the values of “risk”, “population” and “new” are in a same day.

“riski” is the “daily immigration risk” of region i in one day. (i can be 1, 2, 3, …, n. “n” is a fix number. And in this paper “n” is 30, because this paper analyzed 30 regions.)

“populationj” is the daily incoming immigration population from source region j. (j can be 1, 2, 3, …, n, and j cannot be same as i.)

“newj” is the new diagnosed people count in source region j in recent 3 days.

***More details of the design for the formula***

The “recent 3 days” in “newj” is derived from the following considerations: Through testing, it is found that the fewer days used recently, the greater the correlation, and it is judged that the number of long-term confirmed diagnosis is difficult to reflect the current “risk” from source regions. But considering the number of new diagnosed people in single day fluctuates greatly, and the incubation period has a deviation of a few days. Therefore, "recent 3 days" is used here.

On the other hand, it has been also considered the value in the brackets of the above formula, need to be divided by the total population of the source region before calculate cumulative, it means that when the value of newj in two source region is same, the people from the region has smaller total population has greater probability as the virus transmitter moving into the destination region. However, after the correlation analysis of the subsequent models, it was found that the correlation coefficient did not improve, but the value of “offset” and “window” are increased to fit the similar correlation coefficient. It is judged that it is easier to be infected with the virus for migrants than non-migrants. So the above formula was used in the end.

**Calculation of correlation coefficient between “new” and “risk”**

At first, it was tried to calculate the correlation coefficient between “daily new diagnosed” and “daily immigration risk”. The two sets of data are both from January 17th to February 11th, 29 days in total. It was found that the correlation coefficient may not be high enough, such as the value of Jiangsu Province is 0.684 (see Figure 1). There are even almost completely uncorrelated, such as the Heilongjiang Province. The value is only -0.014 (see Figure 2). (Because the value of “daily immigration risk” is a very large value, drawing it with “daily new diagnosed” will make “daily new diagnosed” polyline draw as a very low horizontal line. Therefore, when drawing, the risk coefficient was adjusted, and the maximum “risk” value is set to the maximum “new” value, and the remaining “risk” values are reduced in proportion.)

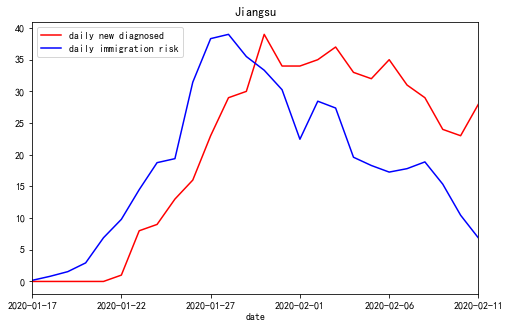


Figure 1. The correlation coefficient of Jiangsu Province is 0.684 from 2020-01-17 to 2020-02-11, when offset=0, window=1

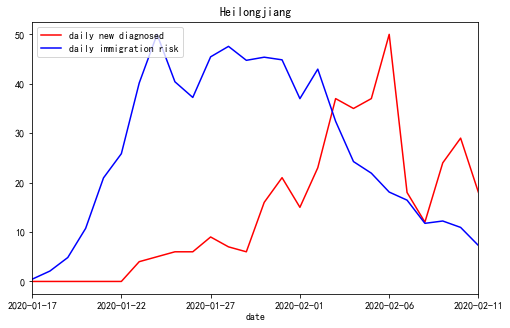


Figure 2. The correlation coefficient of Heilongjiang Province is -0.014 from 2020-01-17 to 2020-02-11, when offset=0, window=1

However, from these two figures, we can see the “new” line and “risk” line has some relationship. Especially figure 1, it is easier to see the shapes of the lines are so similar. So, the parameter “offset” and “window” is introduced as follow.

**Control for incoming immigration population and “offset” parameter**

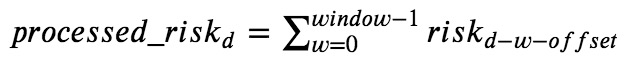
On the one hand, for the incoming immigration population, the preventive measures will be different at different times or in different regions. For example, in some areas, the temperature of immigrants, cough and other medical examinations have been strictly investigated relatively early [6]. Such strict external people flow control will cause infected immigrants are immediately diagnosed, so the number of “daily new diagnosed” will increase significantly on the day when the “risk” factor is high or the day after. On the contrary, these externally infected people will be finally diagnosed relatively later. Therefore, the concept and the parameter of “offset” is introduced: the “risk” is shifted by a certain number of days, for example offset = 3, which means that the “new” of each day is compared with the “risk” of 3 days ago. The correlation coefficient are calculated under offset = 0 in figure 1 and 2.

**Control for inner people and “window” parameter**

On the other hand, the control of internal people and their own awareness of prevention will also affect the spread of the epidemic. For example, infected people who have moved inside the region (have been sick or incubated) and did not seek medical treatment or home isolation, and poor awareness such as going out without a mask will delay the external “risk” until the next few days before the “new” diagnosed people appear. In other words, the regions have the phenomenon of person-to-person transmission. On the contrary, only externally infected people are infected themselves, there will be only “offset” alone. Therefore, the concept of time window (“window” in short) is introduced into the model. For example, if the “window” is 10, it means that the total “risk” for 10 consecutive days will affect “new” on the 10th day. From reference [4], we can know the incubation period with 95% confidence interval is 4.1-7.0 days, so if the infected person who moved in 10 days ago can still affect the inside, it should be much serious from person to person. The correlation coefficient are calculated under window = 1 in figure 1 and 2.

**Use “offset” and “window” to process “risk” before calculate correlation coefficients**

Based on the analysis and the concepts introduced above. The “risk” can be processed by “offset” and “window” as follow formula:



In the formula, all the “processed\_risk” and “risk” is for a same region.

“processed\_riskd” is the value of processed “risk” by “window” and “offset” at the date of “d”.

“riskd-w-offset” is the value of “risk” at the date of “d-w-offset”.

For example, when it is need to calculate the value of “processed\_risk” in 2020-02-10 under window = 2, offset = 3. The formula is as follow:



Now, if we use “offset = 0, window = 9” to get the “processed risk” data for Jiangsu Province, and then calculate the correlation coefficient between “new” and “processed risk”. The value of calculation is increased to 0.979, and the two line charts are significantly more fitted. It is shown as figure 3.

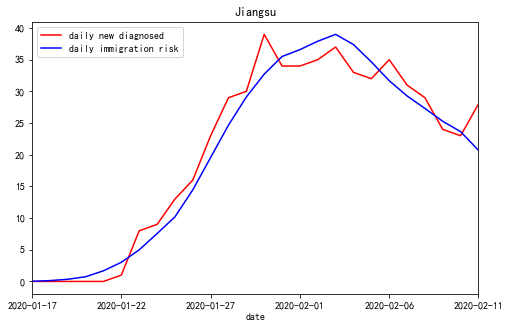


Figure 3. The correlation coefficient of Jiangsu Province is 0.979 from 2020-01-17 to 2020-02-11, when offset=0, window=9

Similarly, use “offset = 4, window = 10” for Heilongjiang Province, the value is increased to 0.874, as shown in figures 4.

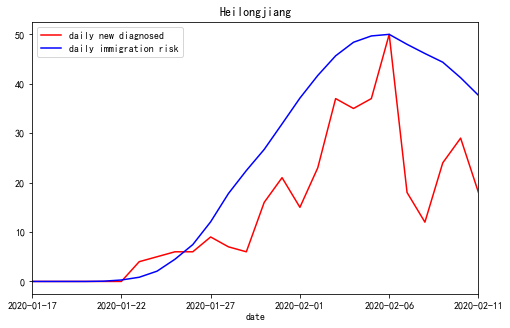


Figure 4. The correlation coefficient of Heilongjiang Province is 0.874 from 2020-01-17 to 2020-02-11 when offset=4, window=10

**Find the best fit for “new” and “risk” by the process with “offset” and “window”**

Now, the final job of this model is to find a set of “offset” and “window” as the best fit for “new” and “processed risk” of each region in each day.

For each region, and for each day, starting from January 17th, 2020, the “offset” is set from 0 to 10, and the “window” is set from 1 to 10. There are 110 different “offsets” and “windows” sets. The 110 sets are used to process “risk” one by one, and calculate the 110 correlation coefficients with “new” and “processed risk”. Finally, the set of “offset” and “window” corresponding to the maximum correlation coefficient is just the model output for the region in the day. The result is show in next section.

**Results**

**Overview**

In this section, the outputs of the model are shown. And the correctness of the model has been confirmed some other data such as R0, news, and etc. Finally, the advantages of evaluation of the effectiveness of epidemic prevention and control by the method of this paper over R0 are analyzed.

**Outputs of model**

Because the result data is a quite large (over 40 days for 30 regions), the representative result data in February 11th for 9 regions, which are compared with related news and data later, is shown in Table 1. (The reason for choosing February 11th is that almost all regions have reduced their daily increase to half of the peak until February 11th or earlier.) If you need full outputs, you can find them from the attachments of this paper.

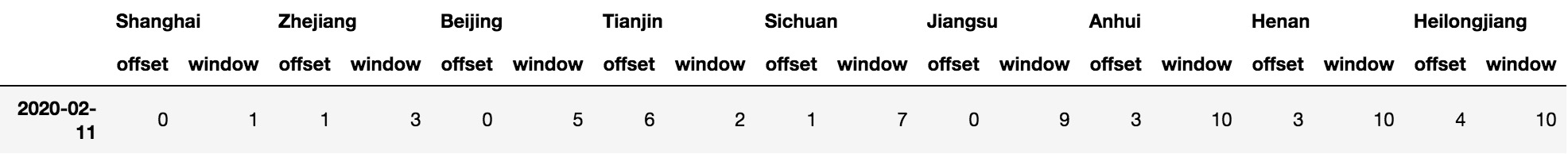


Table 1. “offset” and “window” for 9 regions in February 11th

This table can be used as an evaluation of the effectiveness of prevention and control of the epidemic in each region. The regions are sorted by the values of “offset + window” in ascending order. We can see effectiveness of prevention and control in Shanghai, which owns the lowest “offset” and “window” values, has done the best job in the 9 regions. On the contrarily, the effectiveness in Heilongjiang is worst.

**Related news**

The correctness of the model output is confirmed by some related news as follow.

On the one hand, in late January, there were a large number of infected businessmen who were returning to Wenzhou, Zhejiang from Wuhan, Hubei [5]. On February 1th, the government of Wenzhou, Zhejiang issued 25 strict preventions and control measures in a timely manner [6,7]. On February 22, after the Wenzhou epidemic was completely under control, the Chinese government's newspaper published an article strongly affirming the achievements of Wenzhou's epidemic prevention and control [8]. Although Zhejiang (the province where Wenzhou is located) has the huge pressure, the values of “offset” and “window” are relatively low. So, Zhejiang did a good job.

On the other hand, on February 8th, it is pointed out that Heilongjiang did not pay enough attention to the epidemic and had poor awareness of prevention [9]. Judging from the “offset” and “window” values of Heilongjiang, this is indeed the case.

**Related data**

The correctness of the model output is also confirmed by some related data as follow.

First, according to the diagnosis data released by Doctor Lilac Network in China, the numbers of cumulative diagnosed people are divided into incoming immigrants and inner citizens in only 3 regions: Shanghai, Beijing, and Tianjin. As of the afternoon of February 25th (approximately all outbreaks in these three places are under control), the data is show as table 2. And the model output data near that day is shown as table 3. We can focus “inner diagnosed rate” and “window” values in these two tables. Shanghai did best control (owns lowest “window”) so the inner people diagnosed rate is lowest. Beijing is second, and the rate of Tianjin (owns highest “window”) is the highest.

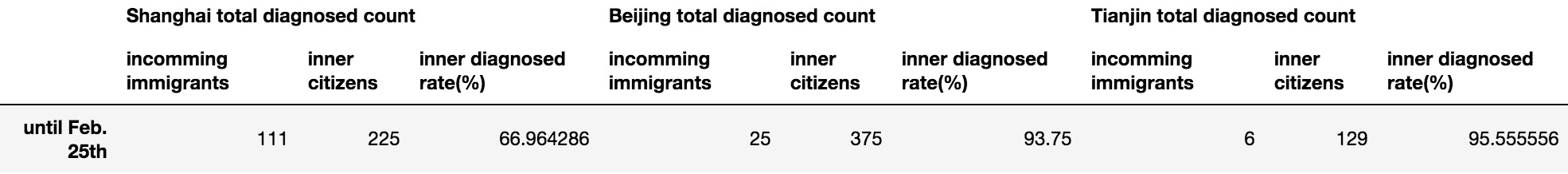


Table 2. The total diagnosed people count divided into incoming immigrants and inner citizens



Table 3. The model outputs for Shanghai, Beijing, Tianjin from February 23th to February 25th

Second, the author of reference [10] collected R0 data from Shanghai, Zhejiang, Jiangsu, Anhui, Henan, and Sichuan (excluding Hubei, Wuhan, and China, which are not involved in this paper).

The R0 values and the “offset + window” values in the date of the article published (February 10th) are shown as table 2. From the table, we can see the relative values and sorting of R0 and “offset + window” is fairly similar.

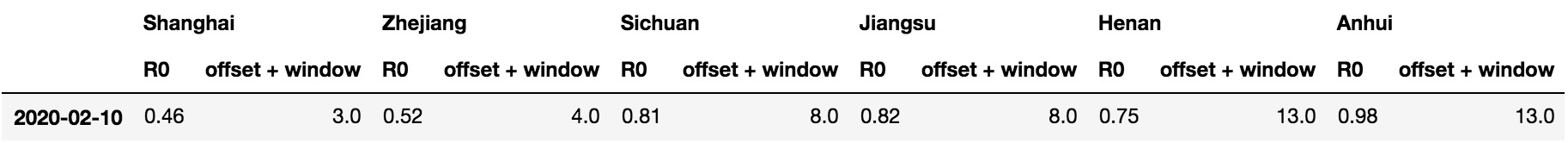


Table 4. The values of R0, “offset” and “window” in February 10th

**The advantages of this model over R0**

Although from Table 4, the evaluation results of the model and the one of R0 are similar, the model has advantages in at least the following two aspects:

First, as described in the previous sections, this model can evaluate by external and internal prevention and control.

Second, this model can find the risk of epidemic prevention and control early. Use the best and worst region in table 1 as the example. Draw lines chart for "new", "offset", "window" and "offset + window" in these two region as figure 5 and 6. We can see “offset + window” of Shanghai is obviously smaller than the ones of Heilongjiang from the dates their fist diagnosed people published. In other words, this model is able to detect some problems before the outbreak.

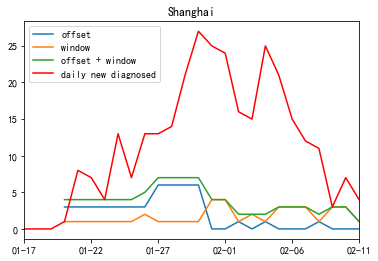


Figure 5. Shanghai’s “new”, “offset”, “window” and “offset + window”

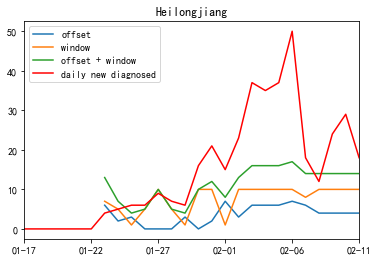


Figure 6. Heilongjiang’s “new”, “offset”, “window” and “offset + window”

In table 5, we can see the values of in their first 2 days when diagnosed people published in detail with all regions mentioned in this paper. The R0 value at the end of January has not got and been compared with the data in the table to prove its correctness, and the amount of data in the previous days is small and the analysis results may be biased by chance. But relatively speaking, the assessment of problem regions in the table is not entirely groundless.

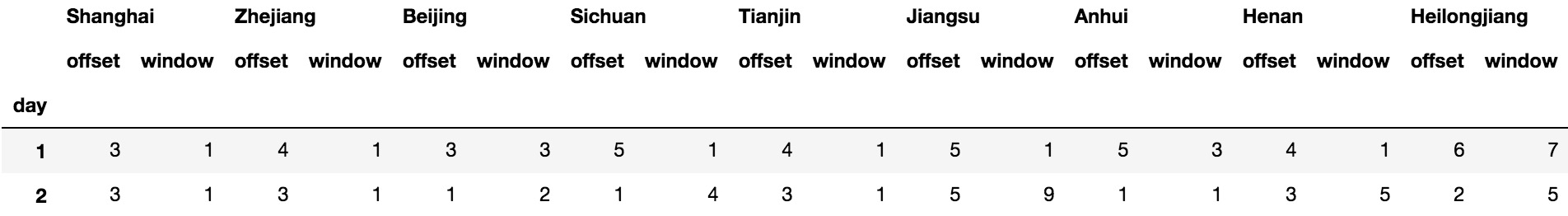


Table 5. All regions’ “offset” and “window” in their first 2 days when diagnosed people published

However, in a certain region, it is impossible for R0 to have calculation results in the first 3 days, because the virus has a high probability of incubation period of at least 3 days. But this model makes it possible. Therefore, this model has the ability to evaluate the effectiveness of prevention and control earlier.

**Discussion**

**Principal Findings**

This study analyzes the relationship of “daily new diagnosed” and “daily incoming immigration population”, and puts forward the concept and calculation method of the “daily immigration risk”. Then, it uses the concept of “offset” and “window” to process the “daily immigration risk” to fit the “new” and “risk” data by maximizing correlation coefficient. And use “offset” and “window” to evaluate the effectiveness of external and internal epidemic prevention and control. At last, this method of evaluation is confirmed by news and data. And, it is more effective than evaluation by R0 values in some respects.

**Limitations**

First, when the value of “daily new diagnosed” drops to 0 or near 0 from the highest peak, the evaluation does not have much reference value.

Second, in the study, it is also found that if the model is used to predict epidemic trends, other aspects need to be considered. For example, Heilongjiang and Jiangxi, although the effectiveness of epidemic prevention and control are not good, but also can completely control the epidemic situation in the middle and late February. On the one hand, there may be improvements in the late few days. However, it is found that another important reason cannot be ignored, that is, the number of incoming immigration population into these two regions and the "risk" value after Spring Festival (January 25th) are much smaller than the ones before January 24th. The decline is much larger than in other regions.

Although the model needs more data to predict the trend of the epidemic, this does not change the fact that if these poorly done regions could do better then, the number of infected people could be reduced. As for the prediction of the epidemic trend mentioned above, it is beyond the scope of this study. At present, I am starting to do the research of trend prediction based on this model.

**Conclusion**

At present, although the epidemic prevention and control in China has achieved great results, it continues. New outbreaks are gradually emerging around the world. The model in this paper may be applicable to the analysis of epidemic prevention and control effectiveness in various parts of the world. I hope this paper can contribute to the prevention and control of the world epidemic. It ’s like a Chinese saying, we, the people all over the world, live in a global village, and we are a family. We Chinese people hope that the epidemic will be brought under control worldwide as soon as possible and finally ended.

**References**

1. CCTV Network: "Banquet for Ten Thousand People" Turns into "Disastrous Banquet", and the history cannot be repeated. Netease News. URL: <https://news.163.com/20/0206/20/F4NTKBK60001899O.html> [accessed 2020-02-06]

2. Notice of the General Office of the State Council on Extending the Spring Festival Holiday of 2020. Chinese Government. <http://www.gov.cn/zhengce/content/2020-01/27/content_5472352.htm>

3. Summary of closure information for tourist attractions and some venues across the country. Happy Chaobai River Community in Sohu.com. URL: <https://www.sohu.com/a/368851082_717402> [accessed 2020-01-25]

4. Qun Li, M.Med., Xuhua Guan, Ph.D., Peng Wu, Ph.D., Xiaoye Wang, M.P.H., Lei Zhou, M.Med., Yeqing Tong, Ph.D., Ruiqi Ren, M.Med., Kathy S.M. Leung, Ph.D., Eric H.Y. Lau, Ph.D., Jessica Y. Wong, Ph.D., Xuesen Xing, Ph.D., Nijuan Xiang, M.Med., et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. The New England Journal of Medicine, January 29th, 2020 DOI: 10.1056/NEJMoa2001316. URL: <https://www.nejm.org/doi/full/10.1056/NEJMoa2001316?query=featured_home> [accessed 2020-01-29]

5. Outside Hubei, the city has the most confirmed diagnoses, with 227 cases! Wuhan is their "second hometown" Jiang You, China Fund News URL: <http://sa.sogou.com/sgsearch/sgs_tc_news.php?req=36lX5Aw6IoHk0uKWsTZiH9M3ZaYRYhjRcv0vr9jcX6Q=> [accessed 2020-01-31]

6. Wenzhou launches comprehensive prevention and control fight. Wenzhou government. <http://www.wenzhou.gov.cn/art/2020/1/29/art_1217829_41860286.html> [accessed 2020-01-29]

7． Wenzhou comprehensively implements "25" emergency measures for epidemic prevention and control. Wenzhou government. <http://www.wenzhou.gov.cn/art/2020/2/1/art_1217828_41865258.html> [accessed 2020-02-01]

8. Wenzhou, Zhejiang: "War" epidemic in the city highlights Wenzhou's role. People's daily. URL: <http://news.cyol.com/content/2020-02/22/content_18390200.htm> [accessed 2020-02-22]

9. Why Heilongjiang has the highest number of confirmed diagnoses in northeast of China. Doctor Lilac Network. URL: <https://mama.dxy.com/outbreak/daily-of-nationwide-new?index=20200208&locationIds=999&share=true&entry=Previous_Review> [accessed 2020-02-08]

10. Qingzhu Dai. When will the epidemic end? The reliable answer is here!. Alibaba Staff Blockchain Alliance. <https://mp.weixin.qq.com/s/6I1onP3jt3G76IYjy9YWlw> [accessed 2020-02-10]