Perspectives

An Estimation of the Total Number of Cases of NCIP (2019-nCoV) — Wuhan, Hubei Province, 2019–2020

Chong You[1]; Qiushi Lin[1]; Xiao-hua Zhou[1,2,#]

Background

On December 31, 2019, the World Health Organization (WHO) was alerted to several cases of pneumonia of unknown etiology in Wuhan, the capital city of Hubei Province in Central China. A novel coronavirus (2019-nCoV) was identified as the causative virus by Chinese authorities on January 7, 2020 (*1*), and China CDC has named the associated disease as novel coronavirus-infected pneumonia (NCIP) (*2*–*3*). As of January 23, 2020, the National Health Commission (NHC) of China had confirmed a total of 830 cases of NCIP in Mainland China, including 177 in critical condition, 25 fatalities, and 34 recoveries. In Wuhan the origin of the NCIP outbreak, 495 cases including 24 fatalities have been confirmed. In addition to the cases in Mainland China, two cases have been detected in Hong Kong Special Administrative Region, China, two in Macao Special Administrative Region, China, one in Taiwan, China, and a total of nine cases have been detected outside China in Thailand (4 cases), Vietnam (2 cases), USA (1 case), Japan (1 case), Republic of Korea (1 case) and Singapore (1 case) (*4*).

The current epidemiological information has indicated that most of the global cases were directly imported from Wuhan. Therefore, a careful and precise understanding of the total number of cases in Wuhan is crucial for decision making and prevention of NCIP. There has already been a considerable investment of resources in Wuhan in order to combat the spread of NCIP. However, estimating the magnitude of the epidemic in Wuhan based on the reported number of confirmed cases is difficult due to the virus' lengthy incubation period and variable symptom presentation (sometimes without the presence of fever or other symptoms), overburdened medical resources and personnel, and time added to receive test results from China's NHC and China CDC. In this article, a method for estimating the total number of NCIP-onset cases within Wuhan is proposed using the number of cases detected outside Hubei Province.

Results

A total of 3,933 cases of NCIP have been estimated in Wuhan (95% confidence interval [CI]: 3,454–4,450) that had an onset of symptoms by January 19, 2020. The estimate, which uses a statistical model (*5*) of 1,723 cases (95% CI: 427–4,471), was given by another research team on January 12, 2020. Compared with that model (*5*), the existing model is improved by 1) including more data from regions outside Wuhan in the model rather than just the three (now nine) confirmed cases outside China used in that model (*5*), which leads to obtaining a much narrower CI and 2) letting the probability of traveling to region i, namely p_i, be different for different i rather than a constant in that model (*5*), which establishes a more realistic and elaborate model.

Assumptions

The proposed model is based on the following assumptions:

1. Wuhan International Airport has a catchment population of 19 million individuals.

2. There is, on average, a d=10-day window between infection and detection, which includes a 5- to 6-day incubation period and a 4- to 5-day delay from symptom onset to detection.

3. Trip durations are long enough that a traveling patient infected in Wuhan will develop symptoms and be detected in other places rather than after returning to Wuhan.

4. All travelers departing from Wuhan, including transfer passengers, have the same risk of infection as local residents.

5. We only consider symptomatic cases with disease severity of a level that can be detected and do not consider asymptomatic or mild cases.

6. Traveling is independent of the exposure risk to

2019-nCoV or of infection status.

7. Patient recoveries are not considered in the model.

8. The proportion of adjusting for the total passengers by air travel volume is a constant over different regions.

Aside from Assumption 8, the same assumptions were also made in the previous model (*5*). Assumption 7 was not explicitly stated in the previous model (*5*) but is implicitly required. Some of the above assumptions are unrealistic, but the data needed to account for these assumptions are not currently available. The following points are further noted:

(i) Violation of Assumption 2 (e.g., the mean time from infection to detection is longer than 10 days) would cause an overestimation of the total number of cases in Wuhan.

(ii) Violation of Assumption 4 (e.g., travelers have a lower risk of infection than residents in Wuhan) would cause an underestimation.

(iii) Violation of Assumption 6 (e.g., infected individuals are less likely to travel due to the health condition) would cause an underestimation.

(iv) Given that there are very few cases of recovery before January 19, 2020, Assumption 8 should not significantly influence the outcome.

Methods

Table 1 lists the top four regions outside of Hubei Province with a relatively large number of reported confirmed cases alongside the corresponding maximum seating capacity for flights from Wuhan. The number of confirmed cases is positively related to passenger volume from Wuhan. Hence, the following model was considered: the number of imported cases $X_{K+d,i}$ from Wuhan to region i by Day (K+d) has a Binomial ($10N_K$, p_i) distribution, i=1, 2, ⋯, m, where N_K is the total number of cases in Wuhan by Day K to be estimated, p_i is the daily probability of traveling from Wuhan to region i, which can be estimated using the ratio of daily volume of passengers and the catchment population of Wuhan airport, and d is the mean time from infection to detection (see details of the model explanation in Appendix A). The calculated daily number of travelers based on flight capacity is further described in Appendix B.

TABLE 1. Number of confirmed cases and seating capacity for 4 regions in China.

Region	Total seats	Cases
Guangdong	111,624	53
Zhejiang	46,528	43
Beijing	59,364	26
Shanghai	51,517	20

Determining the number of imported cases in region i, namely $X_{K+d,i}$, plays a crucial role in the modeling procedure. Table 2 shows the number of reported confirmed cases in various provinces/cities/countries (excluding Hubei Province) within and outside of China on January 23, 2020. The column titled "No. of Local Cases" indicates the number of cases which were not directly imported from Wuhan. Despite the rapid spread of the epidemic, the current situation outside Hubei Province is relatively controlled given the adequate medical support being allocated towards the current outbreak. This suggests that the number of reported cases outside Hubei, as of January 23, 2020, is a fairly accurate representation of the actual epidemic situation in the surrounding regions. Note that only cases directly imported from Wuhan were considered. For example, among the 53 confirmed cases reported in Guangdong Province, of which 8 were local cases, the actual number of imported cases, $X_{K+d,i}$, was regarded as 45. Moreover, for the one case in Singapore, the patient departed from the airport in Guangzhou, hence, it was a non-directly imported case and the corresponding $X_{K+d,i}$ is 0. Furthermore, observations from a few nearby provincial-level administrative divisions (PLADs) including Hunan, Anhui, Henan, Jiangxi, and Tibet and other cities within Hubei Province were dropped due to challenges with estimating daily probability of travel without air transportation data from Wuhan.

Using $X_{K+d,i}$ obtained from domestically and internationally reported cases and the corresponding estimated travel probability p_i, where i=1, 2, ⋯, l, it is possible to infer the magnitude of comparable cases, N_K, within Wuhan that may have occurred on Day K through a binomial model. The MLE estimate of N_K is 3,933 and the corresponding 95% CI is (3,454–4,450). Note that $X_{K+d,i}$ was obtained on 23 January 2020, hence, the estimated N_K is the number of total cases (including those in incubation period) as of January 14, 2020 or the number of cases with symptom onset by January 19, 2020.

Conclusion

The number of confirmed cases in Wuhan reported by China's NHC has increased rapidly in recent weeks.

TABLE 2. Number of reported confirmed cases within (excluding Hubei) and outside China on 23 January 2020.

Region	No. of cases	No. of local cases
Guangdong	53	8
Zhejiang	43	
Beijing	26	
Shanghai	20	1
Chongqing	27	
Sichuan	15	
Guangxi	13	1
Jiangsu	9	
Shandong	9	1
Hainan	8	
Fujian	5	
Tianjin	4	
Liaoning	4	1
Heilongjiang	4	
Jilin	3	
Shaanxi	3	1
Guizhou	3	1
Ningxia	2	
Xinjiang	2	
Gansu	2	1
Yunnan	1	
Inner Mongolia	1	
Shanxi	1	
Qinghai	0	
Hunan	24	1
Anhui	15	
Henan	9	1
Jiangxi	7	
Hebei	2	
Tibet	0	
Macau, China	2	
Hong Kong, China	2	
Taiwan, China	1	
Japan	1	
South Korea	1	
USA	1	
Thailand	3	
Singapore	1*	
Vietnam	2	1

* The patient was a resident from Wuhan city but departed from the airport in Guangzhou.

However, the currently reported number of 495 cases as of January 23, 2020 in Wuhan is still far below our estimate of 3,933. This may be due to the insufficient amount of medical resources in Wuhan and Hubei Province given the suddenness of the outbreak. We suggest boosting medical resources using specific methods such as increasing the amount of hospital beds in order to accommodate all fever patients with pneumonia or a severe respiratory disease in Wuhan in order to expedite the virus examination process and to allow the region to more adequately respond to this public health crisis.

Appendix A

Assume Day 1 is the date of the infection for the very first case. Let N_j denote the number of cases (including those in incubation period) in Wuhan by Day j, Y_j be the number of the cases traveling to region l on Day i, X_j be the number of cases detected in region l by Day j, p is the pre-defined probability of traveling to region l described in Appendix B and d is the mean time from infection to detection (here we suppress the notation l for conciseness). Then Y_j would follow a binomial distribution listed in Table 3 below. Note that from Day d+1 on, the number of trials in the binomial is no longer N_j but N_j-(N_{j-d}-Y_{j-d}) under Assumption 2. Note that Y_{j-d} is relatively small compare with N_{j-d}, hence we drop Y_{j-d} here for simplicity. Therefore,

$$\sum_{j=1}^{K} Y_j \sim Binomial\left(\sum_{j=K-d+1}^{K} N_j, p\right), K > d \qquad (1)$$

However, note that Y_j would not be directly observed on Day j or any other single day but would be detected between a certain period listed in Table 1. For example, suppose that N_K is of interest, then $\sum_{i=1}^{K} Y_i$ needs to be calculated, note that $Y_1,\cdots,Y_K$ would be all included in X_{K+d}, but $\sum_{i=1}^{K} Y_i \le X_{K+d}$ as the observed X_K would include parts of $Y_{K+1},\cdots,Y_{K+d-1}$. A straightforward but rough way to approximate $\sum_{i=1}^{K} Y_i$ is to use $X_{K+d/2}$. The other problem is that using such binomial model, what we can estimate is $\sum_{i=K-d+1}^{K} N_i$ but not a single N_i, we suggest using $\sum_{i=K-d+1}^{K} N_i/\mathrm{d}$ as an estimation of $N_{K-d/2}$, that is

$$X_{K+d/2} \sim Binomial\left(d \times N_{K-d/2}, p\right), \quad K > d \qquad (2)$$

A binomial distribution can be approximated by a Poisson distribution if the number of trials in the binomial distribution is large while the probability of success is small. Hence,

$$X_{K+d} \approx Poisson\left(d \times p \times N_K\right), \quad K > d/2 \qquad (3)$$

TABLE 3. Binomial distributions on Day *i*.

Date	Distribution	Period of Y_i being detected
Day 1	Y_1~Binomial(N_1,p)	Y_1 is expected to be detected on Day d+1
Day 2	Y_2~Binomial(N_2,p)	Y_2 is expected to be detected on Day d+1 and Day d+2
⋮	⋮	⋮
Day d	Y_d~*Binomial*(N_d,p)	Y_d is expected to be detected between Day d+1 and Day 2d
Day d+1	Y_{d+1}~*Binomial*(N_{d+1}−N_1,p)	Y_{d+1} is expected to be detected between Day d+2 and Day 2d+1
⋮	⋮	⋮
Day 2d−1	Y_{2d-1}~*Binomial*(N_{2d-1}−N_{d-1},p)	Y_{2d-1} is expected to be detected between Day 2d and Day 3d−1
Day 2d	Y_{2d}~*Binomial*(N_{2d}-N_d,p)	Y_{2d} is expected to be detected between Day 2d+1 and Day 3d

Including multiple regions into the model, we have

$$X_{K+d,i} \approx Poisson\left(d \times p_i \times N_K\right) \quad \text{for } i = 1, 2, \cdots, m, \qquad (4)$$

and therefore,

$$\sum_{i=1}^{m} X_{K+d,i} \sim Poisson\left(d \times N_K \sum_{i=1}^{m} p_i\right) \qquad (5)$$

where m=25 is the total number of regions used in our model. Note that if p_i=p, our model is almost identical to the previous model (*5*). The total number of cases on Day K, N_K, is estimated by its maximum likelihood estimate (MLE), that is

$$\hat{N}_K = \frac{\sum_{i=1}^{m} X_{K+d,i}}{d \times \sum_{i=1}^{m} P_i} \qquad (6)$$

and the corresponding $(1-\alpha)$ CI is derived using the relation between Poisson distribution and chi-square distribution (*6*).

$$\left(\frac{\chi^2_{2(\sum_{i=1}^{m} X_{K+d,i}),\alpha/2}}{2 \times d \sum_{i=1}^{m} p_i}, \frac{\chi^2_{2(\sum_{i=1}^{m} X_{K+d,i})+2,1-\alpha/2}}{2 \times d \sum_{i=1}^{m} p_i} \right) \qquad (7)$$

Appendix B

The daily probability of traveling from Wuhan to region *i*, p_i, can be estimated using the ratio of daily volume of passengers to region *i* and the catchment population of Wuhan airport. Below are the details for obtaining daily volume of passengers to region *i*.

There were a total of 7,122 flights from Wuhan to 84 airports in Mainland China in the 30 days from December 22, 2019 to January 20, 2020, where 6,586 flights were to the top 50 destinations which accounted for 6,586/7,122=92.47% of the total volume (*7*). Meanwhile, there were 854,383 seats in the flights to top 50 destinations being reported in IATA data in the 22 days between December 30, 2019 and January 20, 2020 (*8*). Hence, the average number of seats in a single flight can be estimated by 854,383/(6,586× 22/30)=177. Over Spring Festival/Lunar New Year, Wuhan airport is expected to handle 24,600 flights and 3.52 million passengers in 40 days (*9*), and thus, each flight is expected to have on average 3,520/24.6= 143 passengers onboard, which gives an average load factor of a flight departing from Wuhan as 143/ 177=0.81. Therefore, the total volume of air travels during the Spring Festival/Lunar New Year can be estimated to be 854,383×0.81/0.9274/22×40=1.35 million. In addition, based on historical evidence, 15 million passengers are expected to depart Wuhan by rail, road, and air, 66% of whom are estimate to travel across 300 km (*10*). That would imply, on average, that 135/(1,500×0.34)=26.47% of trips longer than 300 km would be by air. Therefore, the total passenger volume from Wuhan to other regions in Mainland China can be calculated by *the number of seats*×0.81/0.2647. Note that Hainan Province is a special case because of its geographical location, and a majority of passengers from Wuhan to Hainan Province will likely travel by air. As a result, we would use *the number of seats* ×0.81 for Hainan Province. For other international regions, we use the estimate of 3,301 passengers per day given by the previous model (*5*).

Corresponding author: Xiao-hua Zhou, azhou@math.pku.edu.cn.

[1] Beijing International Center for Mathematical Research, Peking University, Beijing, China; [2] Department of Biostatistics, School of Public Health, Peking University, Beijing, China.

Submitted: January 24, 2020; Accepted: January 28, 2020

References

1. Novel Coronavirus (2019-nCoV). World Health Organization WHO 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019. [2020-1-24].
2. Tan WJ, Zhao X, Ma XJ, Wang WL, Niu PH, Xu WB, et al. A novel coronavirus genome identified in a cluster of pneumonia cases — Wuhan, China 2019–2020. China CDC Weekly 2020;2(4):61–2. http://weekly.chinacdc.cn/en/article/ccdcw/2020/4/61.
3. The 2019-nCoV Outbreak Joint Field Epidemiology Investigation Team, Li Q. An outbreak of NCIP (2019-nCoV) infection in China —

Wuhan, Hubei Province, 2019-2020. China CDC Weekly 2020; 2(5):79–80. http://weekly.chinacdc.cn/en/article/ccdcw/2020/5/79.
4. The current outbreak of novel coronavirus pneumonia on 24 January. National Health Commission of the People's Republic of China. http://www.nhc.gov.cn/yjb/s3578/202001/c5da49c4c5bf4bcfb320ec2036480627.shtml. [2020-1-24] (In Chinese).
5. Natsuko Imai, Ilaria Dorigatti, Anne Cori, Steven Riley, Neil M. Estimating the potential total number of novel coronavirus cases in Wuhan City, China. Imperial College London 2020;1-4. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/2019-nCoV-outbreak-report-17-01-2020.pdf.
6. Ulm K. Simple method to calculate the confidence interval of a standardized mortality ratio. Am J Epidemiol 1990;131(2):373–5. https://doi.org/10.1093/oxfordjournals.aje.a115507.
7. Zhu J. Where might the novel coronavirus has moved before the closure? A forecast based on migration + air + rail data forecast. Urban Data Party https://mp.weixin.qq.com/s/OL_0FZKLFcBXrEvQq28d-A. [2020-1-24] (In Chinese).
8. Where have the residents of Wuhan been in these 22 days of outbreak of novel coronavirus pneumonia? https://mp.weixin.qq.com/s/7ynWYxB-s7nfz7rmjBLSpQ. [2020-1-24] (In Chinese).
9. Wuhan Tianhe airport is expected to handle 3.52 million passengers over the Chinese New Year travel rush. Chutian Metropolis Daily http://www.ctdsb.net/html/2020/0110/hubei284612.html. [2020-1-24] (In Chinese).
10. Big data perspective: Wuhan in the Chinese New Year travel rush. Daily economic news https://m.nbd.com.cn/articles/2020-01-22/1402239.html. [2020-1-24] (In Chinese).