Small Data, Big Time—A retrospect of the first weeks of COVID-19

Qingyuan Zhao[*]

[To be read before The Royal Statistical Society at the Society's 2021 annual conference held in Manchester on Wednesday, September 8th, 2021, the President, Professor Sylvia Richardson, in the Chair]

Abstract

This article reviews some early investigations and research studies in the first weeks of the coronavirus disease 2019 (COVID-19) pandemic from a statistician's perspective. These investigations were based on very small datasets but were momentous in the initial global reactions to the pandemic. The article discusses the initial evidence of high infectiousness of COVID-19 and why that conclusion was not reached faster than in reality. Further reanalyses of some published COVID-19 studies show that the epidemic growth was dramatically underestimated by compartmental models, and the lack of fit could have been clearly identified by simple data visualization. Finally, some lessons for statisticians are discussed.

Keywords: Infectious disease modeling; Selection bias; COVID-19; Model diagnostics.

1 Introduction

Starting from a regional disease outbreak in Wuhan, China, the coronavirus disease 2019 (COVID-19) rapidly grew into a once-in-a-lifetime pandemic. As of the time of writing (December, 2020), the pandemic has already taken at least 1.5 million lives around the world. For anyone living through the COVID-19 pandemic, it is rubbing salt into their wounds to repeat more statistics about the disruptions of lives.

Similar to many other new human diseases, there was an initial period of confusion. On December 31, 2019, the municipal Health Commission of Wuhan first reported a cluster of pneumonia cases of unknown etiology. Because of the exponentially growing nature of a new epidemic outbreak, many researchers raced against time in the first weeks of 2020 to get a better understanding of COVID-19 with very limited data. For the same reason, the investigations in this period had a far-reaching impact and shaped the initial reactions to the pandemic around the world. This article aims to review some of these early investigations from a statistician's perspective.

It is important for the reader to keep in mind that this article is retrospective. Researchers always have limited data at the beginning of a new disease outbreak. This is especially the case for the COVID-19 pandemic: it is fair to say that nobody could have predicted all its rapid and dramatic developments. This article will try to be as objective as possible by presenting the information in a chronological order, although some information known by us now may not have been available to the investigators in the first weeks of the pandemic. Therefore, some degree of hindsight bias is perhaps unavoidable. The main goal of this article is to think about the lessons that can be learned from these investigations to help us respond to future crises. Another goal is to help statisticians to recognize their strengths and prompt others to appreciate the importance of statistics in the era of big data.

This retrospect is also inevitably personal because Wuhan is in fact my hometown. Right after the announcement of the initial cluster of pneumonia cases, I started to follow media reports and research

[*] Statistical Laboratory, University of Cambridge, qyzhao@statslab.cam.ac.uk.

studies. Like many other amateur epidemiologists, I tasked myself with analyzing COVID-19 data and published a preprint in early February to warn that the spread of COVID-19 in Wuhan was much faster than initially thought [25]. After several unsuccessful attempts to publish that article in a scientific journal, I was extremely distressed to watch how the pandemic was quickly developing. Because it is difficult to completely separate facts from speculations (I will try as best as I can), I ask the reader to treat this article as an opinion piece rather than an objective account of the history.

2 Can COVID-19 be transmitted from human to human?

For COVID-19, the most important question before January 20, 2020 was whether it could be transmitted from human to human and thus poses a major threat. To answer this question, the Chinese Center for Disease Control (CCDC) commissioned three groups of expert doctors and epidemiologists in early and mid January. The answer to this question is extremely obvious in hindsight, but that was certainly not the case in early January. By reviewing the key events and evidence in this period, this section will try to understand why it was difficult to conclude that COVID-19 was highly infectious.

2.1 The first warning and some background

Rumors of a novel coronavirus first appeared on Chinese social media on December 30, 2020. The most widespread message was initially posted by Dr. Li Wenliang, who tried to use a group chat to warn fellow doctors who went to the same medical school of a potential disease outbreak [33]. Dr. Li wrote that there had been 7 confirmed SARS (Severe Acute Respiratory Syndrome) cases in the Huanan seafood market and shared the report of a pathogen filtering test showing high confidence of SARS coronavirus, the pathogen responsible for SARS. The message was quickly shared outside the original group. The same evening, two documents from the Wuhan Municipal Health Commission also started to circulate on the internet. Wuhan is the capital city of the Hubei province in China and home to more than ten million people. An official statement was released the next day by the Health Commission, who announced 27 cases of viral pneumonia, many with connections to the Huanan seafood market [45]. The statement was also immediately picked up by a regional office of the World Health Organization (WHO) [18].

Before continuing the story, it may be helpful to go through some background information. Coronavirus is a group of RNA viruses that cause diseases in mammals and birds. Not all coronaviruses are lethal to humans; some strains just cause the common cold. Before the COVID-19 outbreak, two lethal coronaviruses—SARS and MERS (or more precisely, SARS-CoV and MERS-CoV, which are the official names of the viruses)—circulated in humans in this century. The 2002-2003 SARS epidemic infected at least eight thousand people (mostly in China) and caused at least 774 deaths [17]. These figures seem minuscule compared to COVID-19, but SARS indeed had the potential of becoming a major pandemic before being contained in July, 2003 after strict public health interventions. MERS first appeared in 2012 and has not led to any major outbreaks or sustained transmission due to its relatively low reproductive number. However, because of the frequent contact between humans and a natural host of MERS (camels), MERS keeps coming back almost every year. Both SARS-CoV and MERS-CoV are believed to have originated in bats, which also host many other SARS-like coronaviruses [3].

Importantly, it was quickly discovered that the pathogen of COVID-19 is a novel coronavirus that is now known as SARS-CoV-2. As the name suggests, the pathogen is not the original SARS virus but a novel SARS-like coronavirus. So technically speaking, the information provided by Dr. Li Wenliang was not entirely accurate and Dr. Li and several others were reprimanded by the local police for spreading misinformation. As the epidemic unfolded, the reprimand became immensely controversial and was eventually revoked after Dr. Li tragically died from COVID-19 in early February, 2020.

To put the initial reactions to a potentially novel coronavirus in the right context, it is helpful to review the 2002–2003 SARS epidemic briefly. The SARS outbreak first began in the Guangdong province of China in November, 2002, but the pathogen was not identified until mid April, 2003 [19]. The Chinese government was criticized, both domestically and internationally, for its slow responses and delayed communications with the WHO [2]. In fact, the Minister of Health and Beijing's mayor were discharged in April, 2003 after being accused of cover-ups [6]. The course of the epidemic then changed dramatically. After imposing some intensive public health measures (including school closure, travel quarantine, and a purpose-built hospital in Beijing), the SARS epidemic was quickly contained by July.

Some of the initial investigations on COVID-19 were commissioned by the CCDC, which was established in January 2002 and quickly expanded after the SARS epidemic. However, media reports suggest that the CCDC usually only acts as a technical supervisor and lacks administrative authority over the provincial and municipal Health Commissions, which are financed locally [32]. In a landmark project after the SARS epidemic, the CCDC built a comprehensive information system to monitor infectious diseases. In a 2017 thesis that reviews the surveillance system for pneumonia of unknown etiology (PUE), it is concluded that ”the system has played an important role in detection of cases in early stage of human infection with avian influenza outbreak..., but the reporting rate and the positive rate of target pathogen in PUE cases is very low.” Nonetheless, the director of CCDC, Dr. Gao Fu, reassured the public that ”SARS-like viruses can be found at almost any time, but another SARS epidemic will not happen again” in March 2019 [31], just months before the alarm was sounded again.

2.2 The investigations of human-to-human transmissibility

After receiving the first warning, the CCDC immediately sent a group of experts to Wuhan, but the initial investigation did not reach a definitive conclusion. On January 5, 2020, the Wuhan Health Commission announced that the number of pneumonia of unknown etiology cases had increased to 59, but there was ”no clear evidence of human-to-human transmission” [46].

A second group of experts then went to Wuhan on January 8 and reached, surprisingly, a more optimistic conclusion. A statement released on January 11 said that there were no new cases of pneumonia showing symptoms after January 3 and there were no infections among doctors and nurses [47]. One clinician in the expert group, Dr. Wang Guangfa, told the media that the disease ”can be prevented and contained” [39]. In the meantime, the causative pathogen of the pneumonia was officially confirmed as a novel coronavirus on January 9 [38]. The number of suspected cases kept growing rapidly in hospitals across Wuhan [50, 44], but the official case count stagnated at 41 until January 18 (Table 1). This is perhaps why the Wuhan Health Commission maintained the optimistic tone in its statements during this period. Besides reiterating that ”there is no clear evidence of human-to-human transmission”, a January 14 press release further made the following assessment: ”Although we cannot exclude the possibility of limited human-to-human transmission, the risk of sustained transmission is low” [48].

Table 1: Total number of confirmed cases in Wuhan in January, 2020 by date of press release. (Source: Press Releases from the Wuhan & Hubei Municipal Health Commissions.)

Date	12-31	01-03	01-05	01-11	01-12	01-13	01-14	01-15	01-16	01-18	01-19
Cases	27	44	59	41	41	41	41	41	41	45	62
Date	01-20	01-21	01-23	01-24	01-25	01-26	01-27	01-28	01-29	01-30	01-31
Cases	198	258	425	495	572	618	698	1590	1905	2261	2639

However, things took a dramatic turn around January 17. Four new cases in Wuhan were confirmed that day, and more worryingly, the border controls in Thai and Japanese airports had just detected three additional cases who were travelers from Wuhan. By using an estimated number of international travelers from the Wuhan airport and Wuhan's population, a group of researchers in Imperial College

London estimated that a total of 1,723 (95% confidence interval [CI]: 427–4,471) coronavirus cases in Wuhan had onset of symptoms by January 12, which is far more than the official number [5]. Their method was extremely simple, but the results were equally appalling. Moreover, it was discovered on January 15 that in an early family cluster of cases in Shenzhen, one family member had not been to Wuhan in the recent past [34]. This proved that the pathogen could be transmitted between humans. A third expert group was then formed and sent to Wuhan on January 18. Just two days later, it was confirmed for the first time to the public that the new coronavirus was indeed infectious and very likely to be highly infectious. And another three days later, the entire city of Wuhan entered an abrupt and unprecedented lockdown, which was repeated numerous times across the world over the next few months.

2.3 Evidence of high infectiousness

After the lockdown on January 23, 2020, hospitals in Wuhan soon became overwhelmed with patients showing COVID-19 symptoms. The number of deaths also started to skyrocket. In the midst of the emotional outburst after Dr. Li Wenliang's tragic and emblematic death on February 7, the initial investigations came under close scrutiny. Several interviews and articles were published by the Chinese media in the next few weeks and offered a lot of insights into the investigations. Unfortunately, these lessons did not appear to be widely recognized outside China in a quickly developing pandemic. Below I will try to bring us back in time and list some evidence of high infectiousness that was already available by January 10. The reader should keep in mind that, even though the investigative journalism on these investigations appears to be of very high quality, all the interviews and media reports are still retrospective.

First of all, SARS-CoV-2 is not the first coronavirus that has circulated in the human population. Several human coronaviruses produce the generally mild symptoms of the common cold worldwide, and two quite lethal strains (SARS-CoV and MERS-CoV) as mentioned above had led to earlier outbreaks in this century. Being a novel coronavirus, it was quite likely that SARS-CoV-2 was also infectious to humans and poses a major threat to the public health.

Second, although the causative pathogen was officially confirmed as a novel coronavirus on January 9, 2020, there had already been many hints earlier. Weiyuan Gene, a company based in Guangzhou, received a patient sample from the Wuhan Central Hospital on December 24, 2020. Their metagenomic Next Generation Sequencing (mNGS) identified a SARS-like coronavirus in the sample and obtained a near-whole-genome sequence of the virus on December 27 [53]. The sequencing results were shared with the Institute of Pathogen Biology within the Chinese Academy of Medical Sciences. It was reported that a manager at Weiyuan Gene informed the hospital and disease control officials by telephone on the same day the sequencing results came out, and traveled to Wuhan on December 29 to discuss the results [37]. According to another media report, at least nine samples were sent for testing by the end of December. Huada Gene, the industry leader in China, also received a sample and identified a SARS-like coronavirus on December 30 and informed the hospital about the results; the whole genome sequence and the testing reports were then shared with the CCDC and Wuhan Health Commission on January 1 [53]. So it is fair to say that, before the official announcement, the investigators from the CCDC already knew that the causative pathogen was very likely a novel coronavirus.

Third, several hospitals in Wuhan were seeing rapidly increasing numbers of suspected cases in retrospective accounts. In an interview, Dr. Wang Guangfa, a respiratory disease clinician in the second group of experts who went to Wuhan, mentioned that one hospital in Wuhan had a 17% increase in pneumonia cases in December, 2019. By January 10, the 16 Intensive Care Units (ICU) beds at Zhongnan Hospital reserved for pneumonia patients were already filled up [50]. Jinyintan Hospital, a hospital specializing in treating infectious diseases which also became the first designated hospital for COVID-19 in Wuhan, had at least 5 hospital wards full of suspected cases, each with more than 40 patients;

the emergency department at another major hospital was seeing 30-50 patients every day who required immediate hospitalization [41].

In hindsight, there had been enough evidence by January 10, 2020 to conclude that the novel coronavirus was very likely to be highly infectious. So in the perfect scenario, the experts who went to Wuhan could have concluded that the disease outbreak posed a major threat and more stringent public health measures need to be imposed, which in reality only happened after January 20.

2.4 A faster decision?

So why was this conclusion not reached sooner? There are a multitude of reasons. The first is the cautious and conservative approach of the authorities towards infectious diseases. This is wise for most infectious diseases, but perhaps not for a completely novel virus with pandemic potential. On January 3, 2020, all laboratories not affiliated with the CCDC were asked by the National Health Commission on January 3 to destroy their samples and to not disclose their existing results to the public [53], due to the Chinese Law on the Prevention and Treatment of Infectious Diseases [13]. This means that evidence from the previous commercial tests, which had already identified a SARS-like coronavirus a week earlier, could not be relied on in the official investigations. The CCDC received their first patient sample from Wuhan on January 2 [29] and officially announced that the pathogen is a novel coronavirus on January 9 [38].

This conservative approach can also be seen from the case numbers in the press releases of the Wuhan and Hubei Health Commissions (Table 1). The Wuhan Health Commission made no official announcements between January 6 and January 10, but changed the name of the disease from viral pneumonia of unknown etiology to pneumonia due to a novel coronavirus on January 11. Curiously, the case number reported on January 11 dropped from 59 (reported on January 5) to 41; no explanation was specified, but a reasonable guess is that only the samples of 41 patients were tested positive for the new coronavirus. No new cases were announced before January 17, because the local disease control had not received the approved test kits from the CCDC before January 16. During that transition period, a rigorous and lengthy process (that took at least 3-5 days) was needed to confirm a coronavirus case [51], as demonstrated in Figure 1. The final diagnosis, which involves a comprehensive consideration of clinical, epidemiological, and biological evidence, inevitably required a lot of time.

Another factor that delayed the decision is, as reported by several media outlets, strict case definition during the early outbreak. Around January 3, the first CCDC investigators and local experts in Wuhan devised a preliminary guideline in a booklet with a green cover for the treatment of COVID-19 (this was still called viral pneumonia of unknown etiology, or PUE, as the name COVID-19 was designated much later). This booklet followed a 2007 guideline by the Chinese Ministry of Health made after SARS and considered a patient to have a PUE if four clinical criteria were satisfied: fever ($> 38°$C); imaging characteristics of pneumonia; normal or low white blood cells in the early course of disease, or decreased lymphocytes; illness had no improvement or worsened after at least three days of standard antibiotic treatment. [42, 28]. If the suspected case had been exposed to the Huanan seafood market, they only needed to satisfy three of the four criteria above.

What became quite controversial is another booklet with a white cover printed by the Wuhan Health Commission. This ”white booklet” included all the documents in the ”green booklet” and an additional document that makes epidemiological exposure a necessary condition. That is, PUE cases not only needed to show the four clinical symptoms listed above but also had to have a direct epidemiological link to Huanan seafood market [42, 52]. This is a critical mistake in hindsight: the additional requirement made it impossible to conclude that the coronavirus was also quickly circulating outside the Huanan seafood market. It remains unclear who edited the ”white booklet” or why the new inclusion criterion was added, but an anonymous investigator in the first CCDC expert group denied knowledge of the ”white booklet” and the additional document [42]. It is also reported that many doctors on the front line at the time believed that the epidemiological criterion was too strict [50, 42].

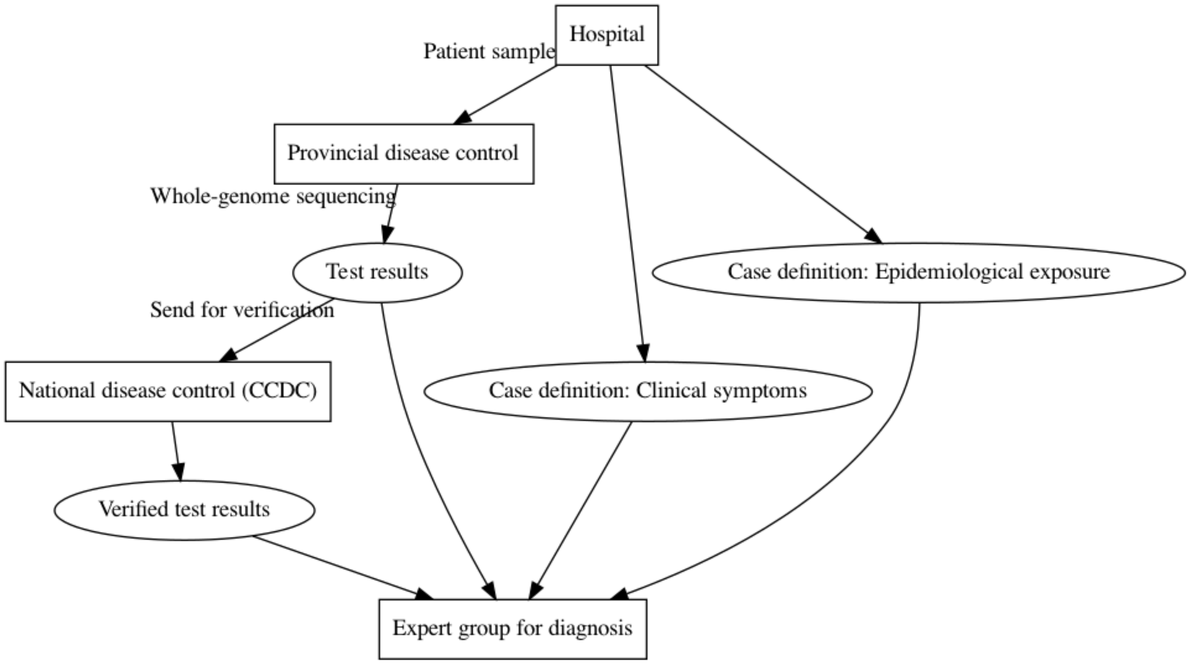
Hospital
Patient sample
Provincial disease control
Whole-genome sequencing
Test results
Case definition: Epidemiological exposure
Send for verification
National disease control (CCDC)
Case definition: Clinical symptoms
Verified test results
Expert group for diagnosis

Figure 1: Case diagnostic process before January 16, 2020 (translated from a diagram in [51]).

A third reason is that the surveillance system designed by the CCDC to prevent another SARS-like epidemic (see Section 2.1) did not work as intended. According to the 2007 guidelines, once a pneumonia case satisfying the definition was found, hospitals were required to organize a consultation (by specialists in the hospital) within 12 hours. If no diagnosis could be reached, the hospital report it as a PUE case through the surveillance system. However, many doctors were unaware of this requirement or the surveillance system [30]. Moreover, some hospitals never reported PUE cases because the patients did not satisfy the epidemiological exposure criterion discussed above [35]. Only a few hospitals started to report their PUE cases through the surveillance system after January 3. These direct reports were then visible to the municipal, provincial, and national Health Commissions [54]. However, the direct reports stopped around January 10 and only resumed in late January. According to media reports, officials at the Wuhan and Hubei Health Commissions asked the hospitals to be extremely cautious about reporting via the surveillance system [43]. From January 4 to January 10, the requirement for reporting through the surveillance system went from ”consultation within the hospital” to ”consultation with experts in the borough”, ”centralized reporting by the city and province”, and finally ”approval by the provincial Health Commission” [54]. So the surveillance system could not provide reliable evidence to infer the severity of the epidemic.

There also appears to have been a lack of statistical expertise and coordination in the first two expert groups. It is reported that the first investigation group (who went to Wuhan on December 31, 2019) included an expert from the CCDC specializing in pathogen identification and two doctors from hospitals in Beijing specializing in respiratory diseases [40]. Although it is certainly important to determine the pathogen of the new disease and understand its clinical features, evaluating the infectiousness of the disease, another crucial question in early disease outbreaks, seemed to be beyond the group’s expertise. The second group of experts included both clinicians and epidemiologists. However, I could not find any media reports suggesting that the epidemiologists tried to gather additional evidence beyond what was officially provided by the local disease control, before concluding that the outbreak ”can be prevented and contained” on January 10, 2020 [39].

Intriguingly, the conclusion that the outbreak could be contained first appeared in an interview with Dr. Wang Guangfa, who is a respiratory disease clinician [39]. In a later interview, Dr. Wang acknowledged that he was not an epidemiologist and the decision was made by the group together [49]. He also mentioned that the epidemiologists from the CCDC could not make a conclusion because there were not enough data to quantify the basic reproductive number R_0 (there were only 41 officially confirmed cases with 2 family clusters). I think this shows a confusion or miscommunication about what "human-to-human transmissibility" means. If it simply means that the disease has the ability to be passed from one human, which is the literal meaning of the phrase in Chinese (and in English), calculating the R_0 is unnecessary and all we need is just one or a few instances like the family cluster in Shenzhen. If it means that R_0 is larger than 1 so a large outbreak will likely happen without interventions, then, due to the lengthy process of case confirmation during that period (Figure 1), the investigation should be focused on the growth of suspected cases or excessive cases of fever and pneumonia. In any case, the public messages sent out by the investigators on January 10 [39] and 11 [47] seem too optimistic and indefensible.

Finally, there appears to be a lack of considerations of risk in the initial investigations, given that SARS already had the potential of becoming a pandemic. In particular, the timing of Wuhan's COVID-19 outbreak was precarious: January 10, 2020 officially marks the beginning of the official travel season (*chunyun* in Chinese) for the Lunar New Year, which in 2020 falls on January 25. Before Wuhan's eventual lockdown on January 23, five million people left Wuhan (typically to go back home or for tourism) [40]. If the new disease was indeed more contagious than SARS, there would be an extremely high risk of spreading the virus across China and to the world. Eventually, this became the main reason behind the unprecedented lockdown of Wuhan on January 23, just two days before the Spring Festival [36]. However, I could not find any articles or interviews discussing risk assessment and management in the initial investigations.

3 How fast was COVID-19 growing?

By late January 2020, it became abundantly clear that the novel coronavirus was infectious and a sizable number of people in Wuhan had already been infected. The next immediate question was: quantitatively speaking, how infectious was the novel coronavirus? How likely was it that the local outbreak would turn into a global pandemic?

Around the same time, I started to read COVID-19 studies and perform some statistical analyses. Prior to this, I had no experience with infectious disease modeling, but I knew a little bit about epidemiology from my research in causal inference. The first thing I noticed was that many studies were trying to estimate the basic reproductive number, or R_0, of COVID-19. The R_0 number is usually defined as the average number of infectees per infected person in the beginning of the outbreak. One can define R_0 more precisely by using a mathematical model for infectious diseases. Something I realized much later is that this also means the precise definition of R_0 is model-dependent.

As some examples, the 2003 SARS epidemic had an estimated R_0 of 2 to 4 [21]; MERS has an R_0 below 1; seasonal strains of influenza usually have an R_0 below 2 [22]. This number is important because if we view the epidemic as a birth process with mean number of offspring equal to R_0, whether R_0 is larger than 1 critically decides the chance of a large outbreak. Moreover, the R_0 value also determines the threshold of "herd immunity" in vaccine development. Besides R_0, another important metric for the infectiousness of a disease is the initial doubling time, which better captures the urgency of the matter.

In this section, I will first briefly review two initial studies on the infectiousness of COVID-19. Because of how early they were published in prestigious medical journals, they attracted massive attention and greatly influenced the initial responses to COVID-19 across the world. Surprisingly, a careful reanalysis of the mathematical models in these studies shows that they do not fit the data in those studies well. I will then discuss alternative ways to analyze early outbreak data and how early studies of COVID-19 understated their uncertainty about R_0.

When reading this section, the reader should keep in mind that, as we now know, there are a significant amount of asymptomatic COVID-19 infections and undetected cases. Under-ascertainment is an especially pronounced issue in the first weeks of the outbreak, and no statistical analysis can completely salvage poor data. Nonetheless, it is still important to apply the most appropriate methods to the available data and make correct interpretations of the results, which many initial COVID-19 studies failed to do.

3.1 Initial studies and mistakes

On January 29, 2020, researchers at the CCDC published the first epidemic incidence curve (by symptom onset date) of COVID-19 on the *New England Journal of Medicine* [8]. This paper analyzed the first 425 confirmed cases in Wuhan and estimated that the initial doubling time was 7.4 days (95% CI 4.2 to 14), the R_0 was 2.2 (95% CI 1.4 to 3.9). These were the first peer-reviewed quantitative estimates of the infectiousness of COVID-19.

In the article, the authors described their statistical analysis as follows:

> We estimated the epidemic growth rate by analyzing data on the cases with illness onset between December 10 and January 4, because we expected the proportion of infections identified would increase soon after the formal announcement of the outbreak in Wuhan on December 31. We fitted a transmission model (formulated with the use of renewal equations) with zoonotic infections to onset dates that were not linked to the Huanan Seafood Wholesale Market, and we used this model to derive the epidemic growth rate, the epidemic doubling time, and the basic reproductive number (R_0).

However, the article provided no particulars about the transmission model and only stated that the model was fitted using MATLAB.

Surprisingly, when fitting a simple Poisson log-linear model to the same epidemic incidence curve, I obtained a quite different estimate: the slope coefficient in the log-linear model, also known as the ”growth exponent” r, is estimated to be 0.187 (95% CI 0.135 to 0.245), which corresponds to a doubling time $\log(2)/r$ of 3.7 days (95% CI 2.8 to 5.1). Figure 2 shows that this simple log-linear model (red curve) provides a good fit with a residual deviance of 27.6 on 24 degrees of freedom, whereas the exponential growth curve corresponding to a doubling time of 7.4 days clearly does not fit the data. Beyond the lack of fit, the restriction to symptom onsets before January 4 was also questionable, as the observed incidences were higher than the predicted values from the log-linear model over the next few days.

If we model an epidemic outbreak as a birth process, the initial growth exponent r can be related to the basic reproductive number R_0 via the identity $R_0 = 1/M(-r)$, where $M(\cdot)$ is the moment generating function of the generation time (time between the infections of successive cases in a chain of transmission) [16]. Estimating the distribution of the generation time is a difficult if not daunting task because the infection times and transmission chains are rarely observed. In practice, it is common to estimate the distribution of serial intervals (time between the symptom onsets of successive cases) using some known transmission chains/clusters and use it to approximate the generation time distribution, but this method is not free from statistical issues [1]. In [8], the authors fitted a Gamma distribution to the serial intervals for just 6 pairs of presumed infector-infectees: 5, 9, 7, 7, 3, 7. They used the estimated serial interval of SARS (mean 8.4 days, standard deviation 3.8 days) as an informative prior and estimated that the serial interval distribution of COVID-19 had a mean of 7.5 days and standard deviation of 3.4 days. No uncertainty quantification was given to these numbers, but they were widely used in subsequent studies.

By using this distribution and the formula for R_0, an initial doubling time of 7.4 days (95% CI 4.2 to 14) would correspond to a R_0 of 2.0 (95% CI 1.4 to 3.2), which is reasonably close to what the article reported. However, if we used the estimated growth exponent from our Poisson log-linear model, the

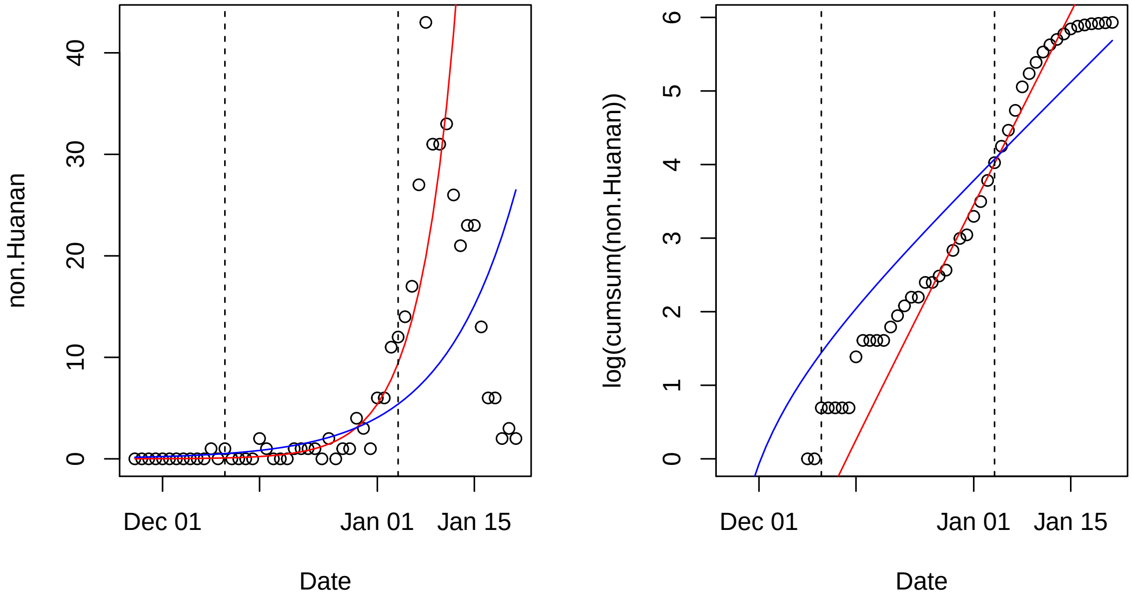

Figure 2: Initial epidemic curve (number of new symptom onsets per day) in Wuhan [8] and the fitted log-linear models. Poisson log-linear models were fitted using the incidences between December 10 and January 4 (dashed lines). The red curves correspond to an unrestricted fit; the blue curves correspond to the best fit assuming that the growth exponent correspond to a doubling time of 7.4 days. The left panel shows the new incidences with no exposure to the Huanan Seafood Market; the right panel shows the logarithm of the cumulative incidences.

corresponding R_0 would be 3.7 (95% CI 2.6 to 5.5). From a policy perspective, these numbers have very different interpretations given that the R_0 of SARS was estimated to be about 2 to 4.

Another influential study modeling the transmission of COVID-19 was published by the *Lancet* on January 31, 2020 [23]. This study used 78 exported cases from Wuhan to areas outside mainland China to fit a susceptible-exposed-infectious-recovered (SEIR) model for the epidemic in Wuhan. To model case exportation, they used a non-homogeneous Poisson process with an estimated volume of outbound air travel from Wuhan. The basic reproductive number is a parameter in their SEIR model and was estimated to be 2.68 (95% CI 2.47 to 2.86); the initial doubling time was estimated to be 6.4 days (95% CI 5.8 to 7.1). A closer read of the article reveals that when fitting this model, the authors only used cases who first showed symptoms before or on January 19, 2020; this end date was chosen ”to minimize the effect of lead time bias on case confirmation”. What the authors did not mention, however, is that this criterion left them with just 17 exported cases.

I fitted a simple Poisson log-linear model to the incidence curve of these 17 cases and the growth exponent was estimated to be 0.117 (95% CI 0.044 to 0.202), which corresponds to a doubling time of 5.9 days (95% CI 3.4 to 15.7). This corresponds to an R_0 of 2.58 (95% CI 1.44 to 4.96), if we follow the *Lancet* study and use the estimated serial interval of SARS (mean 8.4 days, standard deviation 3.8 days). The point estimates I obtained are reasonably close to the original numbers, but the original study seems to have massively understated the variability of the estimates.

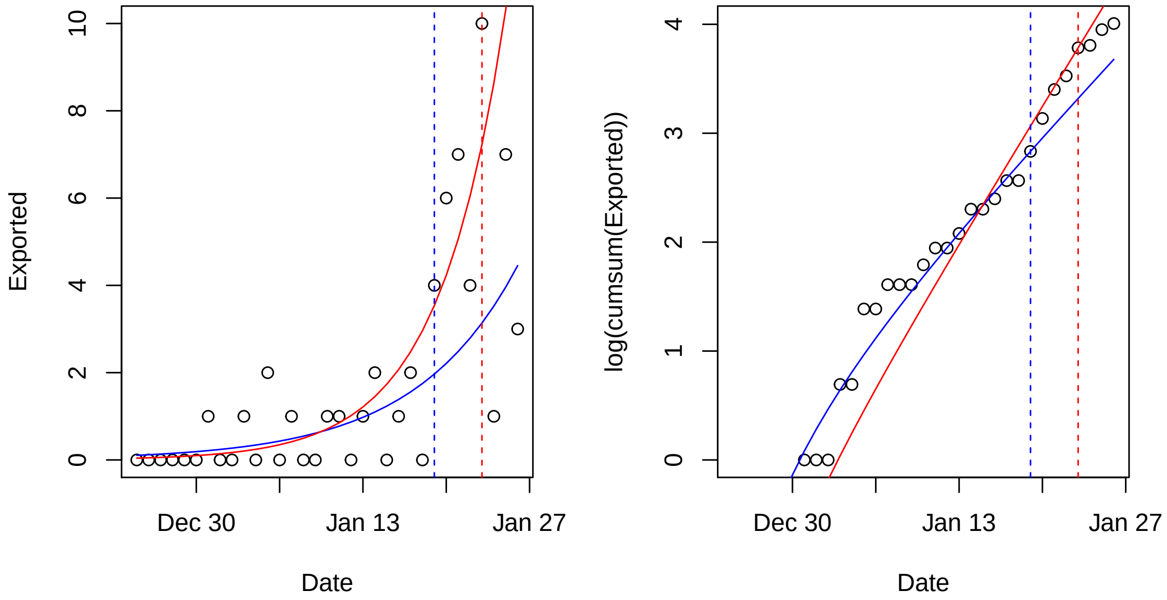

Figure 3: Initial epidemic curve (number of new symptom onsets per day) for Wuhan-exported cases [23] and the fitted log-linear models. The blue curves correspond to fit a Poisson log-linear model using the incidences between December 25 and January 19 (blue dashed lines), the choice in the original study. The red curves correspond to fitting the same model using data up to January 23 (red dashed lines). The left panel shows the new incidences with no exposure to the Huanan Seafood Market; the right panel shows the logarithm of the cumulative incidences.

An examination of the fitted and predicted values shows that this model substantially underestimates the number of exported cases in the few days after January 19; see the blue curves in Figure 3. This shows that the under-ascertainment bias might not have been as large as the authors imagined. If the incidences between January 20 and 23 are also used in model fitting, the initial doubling time would be estimated to be 3.9 days (95% CI 2.9 to 5.5) and the corresponding R_0 would be 4.16 (95% CI 2.77 to 6.60); see the red curves in Figure 3.

Finally, a key missing element in this analysis is Wuhan's lockdown—virtually no civilians were allowed to leave Wuhan after January 23, 2020. This in fact has a nontrivial effect on the incidence curve. For example, consider two cases infected in Wuhan, one on January 15 and one on January 22. It is much more likely for us to observe the first as an exported case than the second, because the second case had a much smaller chance of traveling outside Wuhan before the lockdown.

3.2 Beyond modeling epidemic curves

Next I will explain the basic ideas that are needed to correct the selection bias described in the last paragraph. The illustration below is based on an early study [25] that I was involved in. This study appeared on medRxiv in early February 2020 and was one of the first to suggest that the initial studies substantially underestimated the epidemic growth. Unfortunately, this study is never published.

After Wuhan's lockdown, I started to collect data about the exported cases and immediately noticed that some countries and regions published detailed information about the confirmed cases. As an example, below is an excerpt of the Hong Kong government's press release on January 24, 2020 (translated from Chinese in [26]):

> The other two cases are a married couple of residents of Wuhan, a 62-year-old female and a 63-year-old male, with good prior health conditions. Based on information provided by the patients, they took a high-speed train departing from Wuhan at 2:20pm, January 22, and arrived at the West Kowloon station around 8pm. The female patient had a fever since yesterday with no respiratory symptoms. The male patient started to cough yesterday and had a fever today. They went to the emergency department at the Prince of Wales Hospital yesterday and were admitted to the hospital for treatment in isolation. Currently their health conditions are stable. Respiratory samples of the two patients were tested positive for the novel coronavirus.

The case description clearly contains richer information than the aggregated epidemic curves, although it might not be immediately clear how such information can be useful. In [25, 26], we identified three key events in the trajectory of each case:

1. The beginning of stay in Wuhan, B (the beginning of exposure to the disease);
2. The end of stay in Wuhan, E (the end of exposure to the disease);
3. The onset of symptoms, S.

These events can help in inferring the latent time of infection, T. To understand the relationship between these events, we may model the data using marked point processes (see Figure 4 for two examples). Intuitively, we can use the case description to fill two timelines: a pathophysiological timeline that records the disease progression and recovery (blue in Figure 4), and an epidemiological timeline that records events related to disease transmission (red in Figure 4). Infection is both a pathophysiological and a epidemiology event.

For an exported case, the key event times satisfy the obvious constraint $B \leq T \leq E$. Moreover, $S - T \geq 0$ is the incubation period. Since S is related to the pathophysiological process and B, E are related to the epidemiological process, it may be reasonable to assume that S is conditionally independent of B, E given T (see [26] for more discussion on this assumption). This means that if the distribution of

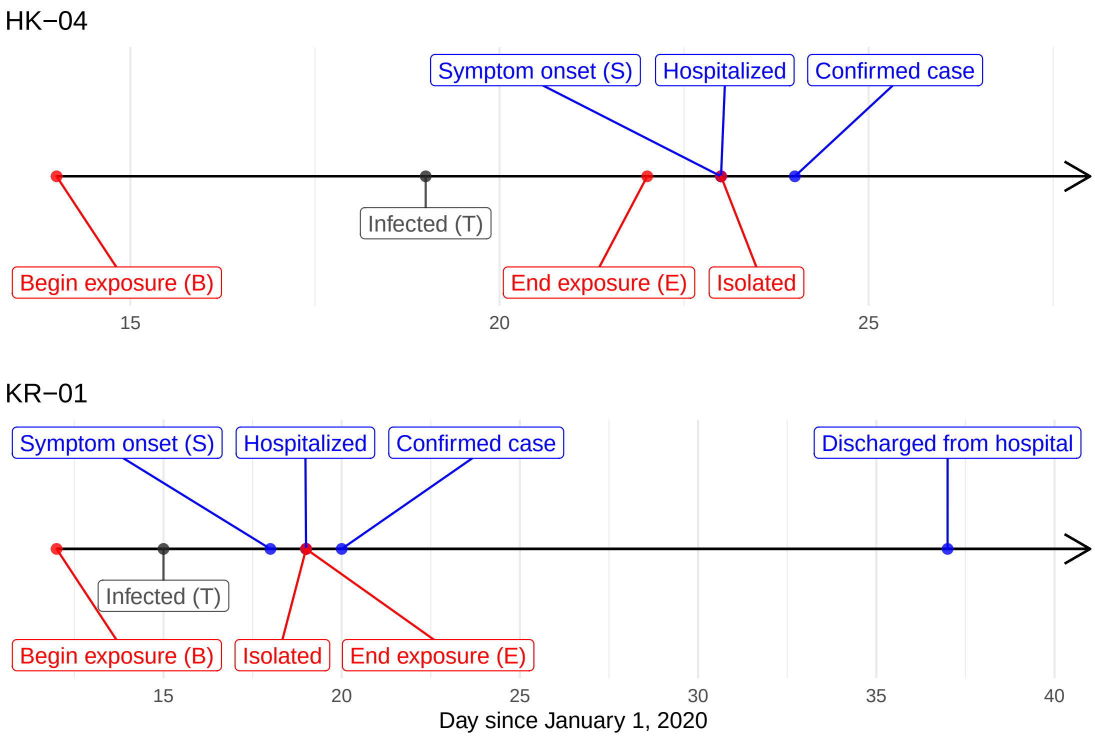

Figure 4: Timelines of two COVID-19 cases [26]. HK-04 is the fourth case in Hong Kong (62-year-old female) and KR-01 is the first case in South Korea (35-year-old female). The color indicates type of event (blue: pathophysiological; red: epidemiological).

the incubation period is known, we can impute the latent infection time T from the symptom onset S by simulating $S - T$ from the incubation period distribution and rejecting the samples that do not satisfy $B \leq T = S - (S - T) \leq E$. We can then fit an appropriate growth model to the imputed T. This is the first basic idea in [25].

Another key observation in [25] is that, because of the travel quarantine of Wuhan, those potential travelers who were infected earlier would have a higher chance of being observed as exported cases. Therefore, instead of using a simple exponential growth model for the density of T

$$f_T(t) \propto e^{rt}, \tag{1}$$

a better model is

$$f_T(t) \propto e^{rt} \cdot \max(L - t, 0), \tag{2}$$

where L is the time of travel quarantine [25]. This model assumes that the probability of observing a exported case infected at time t is decreasing linearly in t until the lockdown; intuitively, this is reasonable if the international travels from Wuhan were stable.

Figure 5 shows a replication of the analysis in [25] with 46 Wuhan-exported international cases collected in [26]; the results are slightly different from the original study because the reanalysis used a slightly different definition of exported cases. The left panel fits the exponential growth model (1) to the expected infection counts (obtained from the imputation algorithm described above) between January 1 and January 20. For the incubation period, I used a log-normal distribution with mean 5.2 and standard deviation 4.9, which is the initial estimate in [8] and the choice in [25]. The estimated growth exponent r is 0.149, which corresponds to a doubling time of 4.6 days. However, this clearly overestimates the number of infections from January 20 to 23 as shown in the figure. The right panel of Figure 5 fits the

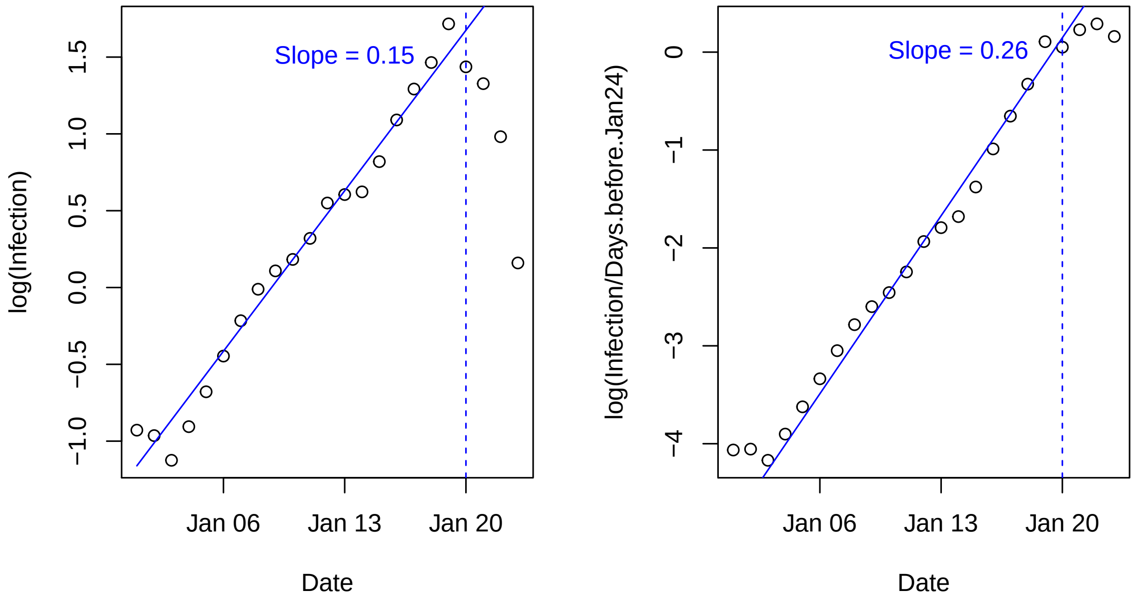

Figure 5: Fitting growth models to the imputed infection time. Data points correspond to the average imputed infection on each day after the transformations specified in the axis label. The left panel fits a simple exponential growth model to the imputed T; The right panel fits the model (2) that is corrected for the observation bias.

bias-corrected model (2) by including an offset. The new estimate of the growth exponent r is 0.260, which corresponds to a doubling time of 2.7 days.

Determining the sampling error of estimates is not trivial. To simplify this replication analysis, I did not use the Bayesian method described in the original analysis [25]. Instead, I fitted quasi-Poisson models corresponding to (2) to 10000 realizations of the transmission times and found that the 2.5% and 97.5% sample quantiles are 0.194 and 0.339. This is not a confidence or credible interval but gives us a sense of the variability of r.

A weakness of this approach is that the incubation period needs to be known. The better method is to jointly estimate the epidemic growth and the incubation period distribution using Wuhan-exported cases. See [26] for more detail, including a theoretical justification of model (2).

3.3 Large uncertainty about R_0

Table 2 collects some early estimates of the initial growth exponent r and basic reproductive number R_0 of COVID-19 as well as the results of the replication analyses above. There is a significant discrepancy between these estimates and some are likely mistaken (Section 3.1). It is worth noting that the two studies that reported the highest R_0 [25, 11] analyzed individual cases, while all the other studies used aggregated incidences.

Most of the epidemiological analyses of COVID-19 estimated the basic reproductive number R_0 instead of the growth exponent r. This is because R_0 can be expressed in terms of the parameters in standard compartmental epidemic models and is more relevant in forecasting the epidemic with various kinds of interventions. But from a statistical perspective, it is much easier to estimate r. Moreover, an initial doubling time of 2 to 3 days, which is what most countries in Europe and elsewhere experienced when the pandemic first hit [9, 26], also highlights the urgency of the matter.

To determine R_0, one needs to estimate not only r but also the distribution of the generation time. As mentioned above, the R_0 and r can be linked through the formula $R = 1/M(-r)$, where $M(\cdot)$ is the moment generating function of the generation time. In practice, it is common to approximate the distribution of generation time by the distribution of serial interval, but this approximation is not very accurate for COVID-19. As an example, the generation time is always positive by definition, but the serial interval can be negative if the infector is asymptomatic. The early estimates of COVID-19's serial interval are also very imprecise due to small sample sizes (the CCDC paper [8] only used 6 pairs of infector-infectees) and methodological issues [1].

It may come as a surprise to statisticians that all the early estimates of the R_0 of COVID-19 in Table 2 did not take into account the uncertainty about the generation time distribution. In fact, had that been considered, many confidence intervals would have become too wide to be useful. To illustrate this, consider the estimated generation time in [7] based on 91 cases in Singapore in February, 2020. In this study, the generation time is modeled by the Gamma distribution and so can be parameterized by its mean and standard deviation. The left panel of Figure 6 shows 100 samples from the posterior distribution of the generation time and the density contours. This plot shows the following generation time distributions are not anomalous in the posterior: mean=4 days, standard deviation=2 days; mean=6 days, standard deviation=4 days. However, they can lead to very different estimates of R_0, as shown in the right panel of Figure 6. In particular, by using the formula $R = 1/M(-r)$, a doubling time of 2.5 days corresponds to a $R_0 = 2.7$ under the first generation time distribution, and a $R_0 = 4.9$ under the second generation time distribution. These R_0 values have drastically different implications for policy decisions, not to mention that this illustration has not taken into account the uncertainty in r yet.

In conclusion, these initial analyses of COVID-19 not only underestimated its R_0 but also underestimated the uncertainty about R_0. They had a far reaching impact on the early global responses to the developing pandemic.

Table 2: Early estimates of r and R_0 of COVID-19.

First author	Published/First posted (in 2020)	Data collected (in 2020)	Sample	Doubling time $log(2)/r$	Reproductive number R_0
Imai [4]	January 25	January 22	7 international exported cases		2.6 (1.5–3.5)
Read [10]	January 28	January 22	Case counts		3.1 (2.4–4.1)
Li [8]	January 29	January 22	425 cases in Wuhan	7.4 (4.2–14)	2.2 (1.4–3.9)
Replication of [8]				3.7 (2.8–5.1)	3.7 (2.6–5.5)
S. Zhao [27]	January 30	January 24	Case counts		Varies with reporting rate: 2.2 (2.0–2.6) to 3.6 (2.9–4.4)
Wu [23]	January 31	January 25	78 international exported cases	6.4 (5.8–7.1)	2.7 (2.6–2.9)
Replication of [23]			Only 17 cases meet the selection criterion	5.9 (3.4–15.7)	2.6 (1.4–5.0)
Q. Zhao [25]	February 9	February 5	46 international exported cases	2.9 (2.0–4.1)	5.7 (3.4–9.2)
Replication of [25]				2.7 (2.0–3.6)	
Yang (later withdrawn) [24]	February 11	January 26	4021 cases in China		3.8 (3.5–4.1)
Sanche [11]	February 11	End of January	140 initial confirmed cases in other Chinese provinces	Method 1: 2.4 (1.9–3.3); Method 2: 2.3 (2.0–2.6)	Method 1: 6.3 (3.3–11.3); Method 2: 6.6 (4.0–10.5)

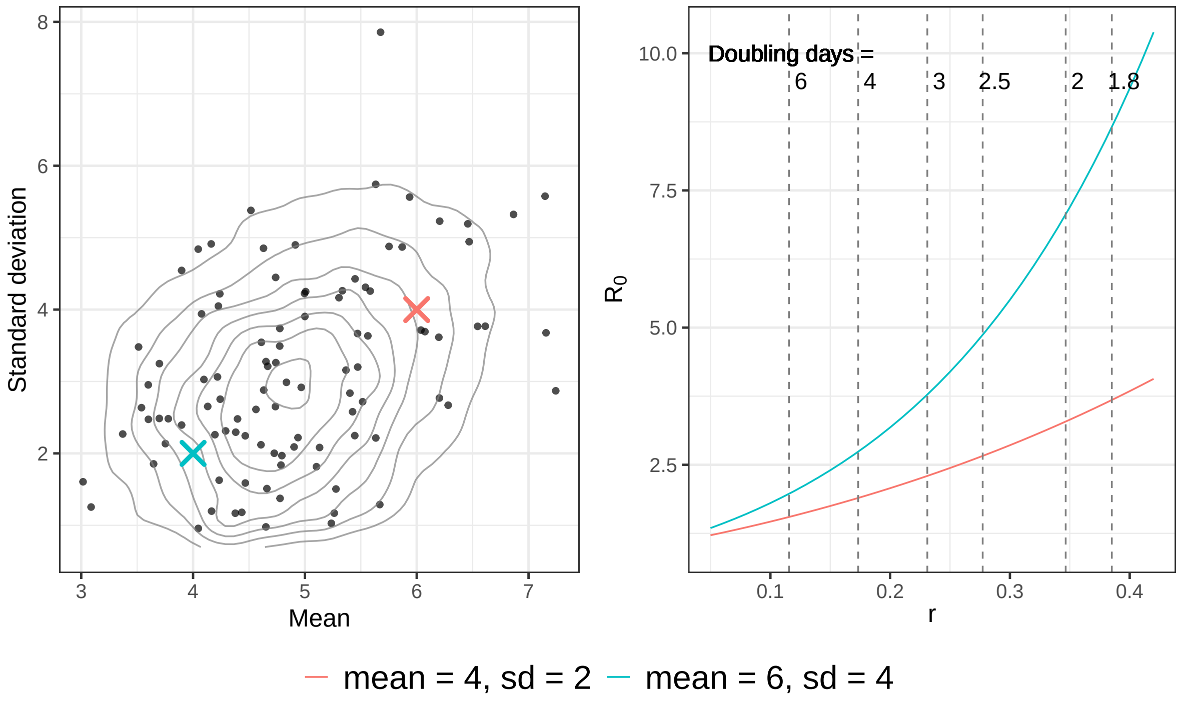

Figure 6: Estimated generation time and the implied basic reproductive number. The left panel shows 100 samples from the posterior distribution of the mean and standard deviation of the generation time in a replication analysis of [7]. The right panel shows the implied R_0 for different values of r and two generation time distributions.

4 Discussion and lessons learned

This article has reviewed some of the most consequential outbreak analyses before COVID-19 turned into one of the worst pandemics in modern history. Reliable data on COVID-19 were scarce, so the initial analyses are understandably imperfect. Nonetheless, had the investigators and decision makers been better prepared, some of the mistakes could have been easily avoided.

In some instances, it is even possible to use common sense to see that the conclusions are likely erroneous. For example, as mentioned above the first CCDC article [8] estimated that the initial doubling time of COVID-19 in Wuhan was 7.4 days. This means that, if community transmission started in mid to late December, 2019, as their data indicate, the epidemic would have gone through only 6 doubling cycles by the time that article was published (January 29, 2020). Consequently, there should have been at most a few hundred cases (as $2^6 = 64$) in Wuhan by late January. However, the number of confirmed cases in Wuhan was already 1,905 on January 29, not to mention that several times more patients with the same clinical features were waiting to be tested. This kind of inconsistency creates confusion in an emerging crisis and can erode the public trust in science and policy.

For statisticians and anyone who engages in data-driven research, I think there are several lessons that can be learned from these mistakes.

Lesson 1: Small data analysis is crucial but not easy

In an emerging disease outbreak, decisions are inevitably made based on limited data. Inference for small samples is a hallmark of statistics and is routinely taught in introductory statistics classes. However, small data analysis is not easy, as exemplified by the initial investigations of COVID-19. The common mistake in many of those investigations is that a general statistical method is applied to new datasets (small or big) without checking if the method is appropriate.

This lesson is of course not new, but its importance in all areas requiring quantitative analysis cannot be understated. In the era of data explosion, many new questions will be asked and will become possible to answer. It will be tempting to apply ”standard” statistical methods to answer these questions, but we must keep in mind that the new data may be collected in non-conventional ways and standard statistical models may be inappropriate.

Lesson 2: Practitioners are still unfamiliar with basic statistical concepts

Another rather discouraging lesson for statisticians is that many applied researchers and the public are still unfamiliar with some of the most basic principles in statistics. Almost no initial COVID-19 studies listed in Table 2 performed any model checking or published their code for reproducibility. Many studies and forecasting models relied on results from other studies, but they rarely considered uncertainty in the prior estimates. Almost all media coverage on the COVID-19 studies only reported point estimates and did not mention any form of uncertainty quantification. Finally, there were almost no serious considerations of selection bias in the initial COVID-19 studies.

Lesson 3: Better research appraisal is in urgent need

COVID-19 has presented a serious challenge to scientific journals and there were some encouraging moves. For example, many notable publishers, journals, and institutions signed a statement on January 31, 2020 to make all publications related to COVID-19 immediately open access [14]. However, journals were quickly overwhelmed with submissions related to COVID-19 and the peer reviews were very unreliable. The high volume of research publications on COVID-19, often showing inconsistent results, further caused confusion. In the end, extra effort is often needed to appraise and synthesize the research studies.

Statisticians can contribute to improving this process in several ways. First, statisticians need to engage in applied research more actively and disseminate the basic statistical principles. Second, statisticians can take a bigger role in peer reviews and post-publication discussion of scientific studies. Many academic journals allow the manuscript type of ”letter to the editor” to discuss published articles and some journals also allow and moderate online comments. However, they are rarely used by or even known to statisticians. Finally, better methods for research synthesis are needed to take data quality and selection bias into account.

Lesson 4: Right data are often more important than right analysis

I would like to close this retrospect by going back further into history. In statistics education, it is common to assume that high quality data are already available and are just ”waiting to be analyzed”. But this is rarely the case in practice. The modern discipline of epidemiology is often traced back to John Snow’s inspiring investigation of the London cholera epidemic in 1854. He refuted the then popular miasma theory and identified that the source of the outbreak was a public water pump on Broad Street in London. From a statistical perspective, Snow’s most innovative idea was to focus on death rates in districts served by two water companies that drew water from different stretches of the Thames River. The huge difference in death rates due to cholera provided the convincing evidence that germ-contaminated water was the cause of the epidemic [20]. By any standard, this is a remarkable achievement because Snow, unlike us, had no access to all the tools offered by modern microbiology and statistics.

Interestingly, a group of modern researchers recovered historical archives in Ferrara, Italy, which also suffered from another cholera epidemic in 1855 [12]. They analyzed the data in these archives using modern statistical methods and found that ”the better kept houses in the better parts of the town had less cholera morbidity and especially mortality", supporting a miasma theory. As pointed out by a discussant of the paper, this is not entirely surprising: the Ferrara data were recorded under the misguided miasmas theory and can never reveal the superiority of the competing germ theory as in Snow's analysis [15].

The same thing can be said about COVID-19. In the initial investigations of human-to-human transmissibility, the main obstacle is a lengthy process of reporting and testing suspected cases (Section 2.2). Although a due process with convincing proofs is certainly needed to confirm transmissibility, more stringent public health interventions could have been imposed sooner if the right data were collected, as argued in Section 2.3. This teaches us another lesson in the era of data: statisticians should not be complacent with being the best analysts of "big data"; instead, we need to step up and play a bigger role in study design and data collection.

Afterword

The above article was initially written in December, 2020. Back then, the United Kingdom, where I live, was heading to a third national lockdown. This article was then revised in June, 2021 and prospects of the pandemic became a lot less grim due to successful vaccines. But unfortunately, the origins of COVID-19 continue to be politicized. When writing the article, it was difficult for me to not imagine a counterfactual scenario in which Wuhan was locked down a week earlier, even though I think the factual lockdown was already an incredibly resolute decision. Could COVID-19 be contained before turning into a pandemic? Could it buy more time for producing personal protective equipment and vaccine development? There are many more hypothetical questions one could ask, but the answers will never be known. Eventually, I reconciled myself through the history of infectious diseases and the lens of evolutionary biology. New kinds of viruses will eventually appear and circulate, and humanity will face greater challenges and need greater solidarity. What is important is not to find fault with others, but to learn from each others' failures.

Acknowledgement

Qingyuan Zhao is supported in part by an early career grant from the Isaac Newton Trust. The author is grateful to the Discussion Papers Editor, Dr. Adam Sykulski, and an anonymous referee for carefully reading a previous version of the manuscript and providing many helpful suggestions.

References

[1] S. Bacallado, Q. Zhao, and N. Ju. Letter to the editor: Generation interval for covid-19 based on symptom onset data. *Eurosurveillance*, 25(29), 2020. doi: 10.2807/1560-7917.es.2020.25.29.2001381. URL `https://doi.org/10.2807/1560-7917.es.2020.25.29.2001381`.

[2] T. Crampton. W.H.O. criticizes China over handling of mystery disease. *The New York Times*, April 7 2003. URL `https://www.nytimes.com/2003/04/07/international/asia/who-criticizes-china-over-handling-of-mystery-disease.html`. Accessed 2020-10-22.

[3] X.-Y. Ge, J.-L. Li, X.-L. Yang, A. A. Chmura, G. Zhu, J. H. Epstein, J. K. Mazet, B. Hu, W. Zhang, C. Peng, Y.-J. Zhang, C.-M. Luo, B. Tan, N. Wang, Y. Zhu, G. Crameri, S.-Y. Zhang, L.-F. Wang, P. Daszak, and Z.-L. Shi. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*, 503(7477):535–538, 2013. doi: 10.1038/nature12711. URL `https://doi.org/10.1038/nature12711`.

[4] N. Imai, A. Cori, I. Dorigatti, M. Baguelin, C. A. Donnelly, S. Riley, and N. M. Ferguson. Report 3 - transmissibility of 2019-ncov. Technical report, MRC Centre for Global Infectious Disease Analysis, January 25 2020.

[5] N. Imai, I. Dorigatti, A. Cori, S. Riley, and N. M. Ferguson. Report 1 - estimating the potential total number of novel coronavirus (2019-ncov) cases in wuhan city, china. Technical report, MRC Centre for Global Infectious Disease Analysis, January 17 2020.

[6] S. Jakes. Beijing's sars attack. *TIME*, April 8 2003. URL `http://content.time.com/time/magazine/article/0,9171,441615,00.html`. Accessed 2020-10-22.

[7] C. Kremer, T. Ganyani, D. Chen, A. Torneri, C. Faes, J. Wallinga, and N. Hens. Authors' response: Estimating the generation interval for covid-19 based on symptom onset data. *Eurosurveillance*, 25(29), 2020. doi: 10.2807/1560-7917.es.2020.25.29.2001269. URL https://doi.org/10.2807/1560-7917.es.2020.25.29.2001269.

[8] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 2020.

[9] L. Pellis, F. Scarabel, H. B. Stage, C. E. Overton, L. H. K. Chappell, K. A. Lythgoe, E. Fearon, E. Bennett, J. Curran-Sebastian, R. Das, M. Fyles, H. Lewkowicz, X. Pang, B. Vekaria, L. Webb, T. A. House, and I. Hall. Challenges in control of covid-19: short doubling time and long delay to effect of interventions. *medRxiv*, 2020. doi: 10.1101/2020.04.12.20059972. URL https://www.medrxiv.org/content/early/2020/06/11/2020.04.12.20059972.

[10] J. M. Read, J. R. Bridgen, D. A. Cummings, A. Ho, and C. P. Jewell. Novel coronavirus 2019-ncov: early estimation of epidemiological parameters and epidemic predictions. *medRxiv*, 2020. doi: 10.1101/2020.01.23.20018549. URL https://www.medrxiv.org/content/early/2020/01/28/2020.01.23.20018549.

[11] S. Sanche, Y. T. Lin, C. Xu, E. Romero-Severson, N. Hengartner, and R. Ke. The novel coronavirus, 2019-ncov, is highly contagious and more infectious than initially estimated. *medRxiv*, 2020. doi: 10.1101/2020.02.07.20021154. URL https://doi.org/10.1101/2020.02.07.20021154.

[12] C. Scapoli, E. Guidi, L. Angelini, A. Stefanati, and P. Gregorio. Sociomedical indicators in the cholera epidemic in ferrara of 1855. *European Journal of Epidemiology*, 18(7):617–622, 2002. doi: 10.1023/a:1024829328204. URL https://doi.org/10.1023/a:1024829328204.

[13] Standing Committee of the National People's Congress. Law of the people's republic of china on prevention and treatment of infectious diseases (2013 amendment). http://www.gov.cn/banshi/2005-08/01/content_19023.htm; English translation in http://www.lawinfochina.com/display.aspx?id=14881&lib=law, 2004. Accessed 2020-11-28.

[14] W. Trust. Sharing research data and findings relevant to the novel coronavirus (covid-19) outbreak. https://wellcome.org/coronavirus-covid-19/open-data. Accessed 2020-12-10.

[15] J. P. Vandenbroucke. Commentary: The 1855 cholera epidemic in ferrara: Lessons from old data reanalysed with modern means. *European Journal of Epidemiology*, 18(7):595–598, 2002. doi: 10.1023/a:1024984102096. URL https://doi.org/10.1023/a:1024984102096.

[16] J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604, 2006. doi: 10.1098/rspb.2006.3754. URL https://doi.org/10.1098/rspb.2006.3754.

[17] WHO. Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. https://www.who.int/csr/sars/country/table2004_04_21/en/, . Accessed 2020-10-21.

[18] WHO. Timeline of WHO's response to COVID-19. https://www.who.int/news/item/29-06-2020-covidtimeline, . Accessed 2020-10-21.

[19] WHO. Coronavirus never before seen in humans is the cause of sars. https://www.who.int/mediacentre/news/releases/2003/pr31/en/, April 16 2003. Accessed 2020-10-21.

[20] W. Winkelstein. A new perspective on john snow's communicable disease theory. *American Journal of Epidemiology*, 142(Supplement 9):S3–S9, 1995. doi: 10.1093/aje/142.supplement_9.s3. URL https://doi.org/10.1093/aje/142.supplement_9.s3.

[21] World Health Organization. Consensus document on the epidemiology of severe acute respiratory syndrome (sars). Technical report, World Health Organization, 2003.

[22] World Health Organization. Who public health research agenda for influenza: minimizing the impact of pandemic, zoonotic, and seasonal epidemic influenza. Technical report, World Health Organization, 2017.

[23] J. T. Wu, K. Leung, and G. M. Leung. Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. *The Lancet*, 395(10225):689–697, 2020. doi: 10.1016/s0140-6736(20)30260-9. URL https://doi.org/10.1016/s0140-6736(20)30260-9.

[24] Y. Yang, Q.-B. Lu, M.-J. Liu, Y.-X. Wang, A.-R. Zhang, N. Jalali, N. E. Dean, I. Longini, M. E. Halloran, B. Xu, X.-A. Zhang, L.-P. Wang, W. Liu, and L.-Q. Fang. Epidemiological and clinical features of the 2019 novel coronavirus outbreak in china. *medRxiv*, 2020. doi: 10.1101/2020.02.10.20021675. URL https://doi.org/10.1101/2020.02.10.20021675.

[25] Q. Zhao, Y. Chen, and D. S. Small. Analysis of the epidemic growth of the early 2019-nCoV outbreak using internationally confirmed cases. *medRxiv*, 2020. doi: 10.1101/2020.02.06.20020941. URL https://www.medrxiv.org/content/early/2020/02/09/2020.02.06.20020941.

[26] Q. Zhao, N. Ju, S. Bacallado, and R. D. Shah. BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic. *Annals of Applied Statistics*, 2020. In press.

[27] S. Zhao, Q. Lin, J. Ran, S. S. Musa, G. Yang, W. Wang, Y. Lou, D. Gao, L. Yang, D. He, and M. H. Wang. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-ncov) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases*, 92:214–217, 2020. doi: 10.1016/j.ijid.2020.01.050. URL https://doi.org/10.1016/j.ijid.2020.01.050.

[28] 中国卫生健康委员会. 全国不明原因肺炎病例监测、排查和管理方案. http://www.nhc.gov.cn/bgt/pw10708/200708/4455f46a2f5e4908a8561c079ecbcf0e.shtml, 2007. Accessed 2020-11-28. [English translation: Chinese Health Commission. National surveillance, investigation and management plan for pneumonia cases of unknown etiology.].

[29] 中国疾病预防控制中心. 在一线, 疾控勇士与新型冠状病毒赛跑. http://www.chinacdc.cn/yw_9324/202002/t20200201_212137.html, February 1 2020. Accessed 2020-11-28. [English translation: Chinese Centre for Disease Control. On the front line, Disease Control warriors race against the new coronavirus.].

[30] 信娜, 王小, 孙爱民, and 辛颖. 投资 7.3 亿的传染病网络直报系统因何失灵 28 天. 财经 https://mp.weixin.qq.com/s?__biz=MzA5NTM4Njk2Mw==&mid=2665968779&idx=1&sn=3fdbd7a3736f53d81a139b4a15bb27a7&scene=21, February 25 2020. Accessed 2020-12-02. [English translation: Xin, Na and Wang, Xiao and Sun, Aimin and Xin, Yin. Why the direct reporting system for infectious diseases with an investment of RMB 730 million failed for 28 days? Caijing.].

[31] 倪伟. 疾控中心主任高福：理直气壮告诉大家，中国疫苗很好！. 新京报, March 4 2020. URL http://www.bjnews.com.cn/feature/2019/03/04/552608.html. Accessed 2020-11-26. [English translation: Ni, Wei. Gao Fu, Director of the Centers for Disease Control and Prevention: Tell everyone with confidence that China's vaccines are very good! The Beijing News.].

[32] 八点健闻. 名实不副地位太低，你可能不知道的中国疾控往事. https://www.tmtpost.com/4259786.html, February 28 2020. Accessed 2020-11-26. [English translation: Status does not match its name, things you do not know about the past of Chinese Center for Disease Control. Badian Health News.].

[33] 刘名洋. 对话"传谣"被训诫医生：我是在提醒大家注意防范. 新京报, January 31 2020. URL http://www.bjnews.com.cn/feature/2020/01/31/682076.html. Accessed 2020-10-21.

[34] 南方都市报. 详情披露！广东首个"发病"家庭确诊过程公开. https://wemp.app/posts/c483336b-1a74-49eb-a447-4321400a28c8, February 25 2020. Accessed 2020-11-26. [English translation: Details disclosure! The process of diagnosing Guangdong's first "morbid" family is made public. Southern Metropolis Daily.].

[35] 吴靖. 失去的机会，新冠疫情早期被忽视的小医院病例. 八点健闻 https://new.qq.com/rain/a/20200211A05AEA00, February 11 2020. Accessed 2020-12-01. [English translation: Wu, Jing. Lost opportunities: Neglected cases in small hospitals during the early coronavirus outbreak. Badian Health News.].

[36] 孙梦. 解封在即, 李兰娟首次披露武汉封城细节. 中国卫生杂志 https://mp.weixin.qq.com/s/1MQDiyfiJt9YA9lK5b_i-A, March 26 2020. Accessed 2020-12-02. [English translation: Sun, Meng. Soon before easing the lockdown, Dr Li Lanjuan for the first time discloses details of Wuhan's lockdown. Chinese Health Magazine.].

[37] 小山狗. 记录一下首次发现新型冠状病毒的经历. https://project-gutenberg.github.io/nCovMemory-Web/post/a544e7b4ac5d5ccf01e8259a70b95044, January 28 2020. Accessed 2020-11-27 (the original post has been removed). [English translation: Xiaoshangou (online id). Record the experience of discovering the new coronavirus for the first time.].

[38] 屈婷. 专家称系新型冠状病毒武汉不明原因的病毒性肺炎疫情病原学鉴定取得初步进展. 新华网 http://www.xinhuanet.com/2020-01/09/c_1125438971.htm, January 9 2020. Accessed 2020-11-28. [English translation: Qu, Ting. Experts say it is a new type of coronavirus, and preliminary progress has been made in the etiological identification of the unexplained viral pneumonia epidemic in Wuhan. Xinhua News Agency.].

[39] 廖君 and 黎昌政. 专家称武汉不明原因的病毒性肺炎可防可控. http://www.xinhuanet.com/local/2020-01/11/c_1125448549.htm, January 11 2020. Accessed 2020-11-26. [English translation: Liao, Jun and Li, Changzheng. Experts say Wuhan's unexplained viral pneumonia is preventable and controllable. Xinhua News Agency.].

[40] 李想俣, 李明子, 彭丹妮, and 杜玮. 武汉之憾：黄金防控期是如何错过的？. 中国新闻周刊, 934, February 10 2020. Accessed 2020-12-02 (the original web link has been removed). [English translation: Li, Xiangyu and Li, Mingzi and Peng, Danni and Du, Wei. Regret of Wuhan: How was the golden period to contain the epidemic missed? China News Weekly.].

[41] 杨楠. "重组"金银潭: 疫情暴风眼的秘密. 南方周末, https://www.infzm.com/contents/178385, March 5 2020. Accessed 2020-11-28. [English translation: Yang, Nan. "Reorganization" Jinyintan: Secrets in the eye of the epidemic. Southern Weekly.].

[42] 杨海. 白皮手册与绿皮手册: 新冠肺炎诊断标准之变. 中国青年报-冰点周刊, February 20 2020. URL https://mp.weixin.qq.com/s/vysNta8IU2wbRBv-c3aS4Q. Accessed 2020-11-28. [English translation: Yang, Hai. The White Handbook and The Green Handbook: Changes in diagnostic criteria for the novel coronavirus pneumonia. Chinese Youth Daily–Freezing Point.].

[43] 杨海. 武汉早期疫情上报为何一度中断. 中国青年报-冰点周刊, March 5 2020. URL https://xw.qq.com/cmsid/20200305A0O3MV00. Accessed 2020-12-01. [English translation: Yang, Hai. Why was the case report once suspended during the early outbreak in Wuhan? Chinese Youth Daily–Freezing Point.].

[44] 杨海. 武汉市中心医院医生：传染病留给大家反应的时间太短了. 中国青年报-冰点周刊, March 13 2020. URL https://terminus2049.github.io/archive/2020/03/13/bing-dian.html. Accessed 2020-11-26. [English translation: Yang, Hai. Doctor from the Wuhan Central Hospital: The infectious disease leaves us too little time to respond. Chinese Youth Daily–Freezing Point.].

[45] 武汉市卫生健康委员会. 武汉市卫健委关于当前我市肺炎疫情的情况通报. http://wjw.wuhan.gov.cn/xwzx_28/gsgg/202004/t20200430_1199576.shtml, December 31 2019. Accessed 2020-10-21.

[46] 武汉市卫生健康委员会. 武汉市卫生健康委员会关于不明原因的病毒性肺炎情况通报. http://wjw.wuhan.gov.cn/xwzx_28/gsgg/202004/t20200430_1199589.shtml, January 5 2020. Accessed 2020-11-26. [English translation: Wuhan Health Commission. Situation report on viral pneumonia of unknown etiology.].

[47] 武汉市卫生健康委员会．专家解读不明原因的病毒性肺炎最新通报．`http://wjw.wuhan.gov.cn/xwzx_28/gsgg/202004/t20200430_1199592.shtml`, January 11 2020. Accessed 2020-11-26. [English translation: Wuhan Health Commission. Experts explain the latest situation report on viral pneumonia of unknown etiology.].

[48] 武汉市卫生健康委员会．新型冠状病毒感染的肺炎疫情知识问答．`http://wjw.wuhan.gov.cn/xwzx_28/gsgg/202004/t20200430_1199594.shtml`, January 15 2020. Accessed 2020-11-26. [English translation: Wuhan Health Commission. Q&A about the novel coronavirus pneumonia epidemic.].

[49] 秦珍子．卫健委专家组成员王广发出院了，回答了我们 8 个问题．中国青年报-冰点周刊, February 2 2020. URL `https://mp.weixin.qq.com/s/DWcRVz10zps27VIrml_Khg`. Accessed 2020-12-02. [English translation: Qin, Zhenzi. After being discharged from hospital, Dr Wang Guangfa from Health Commission's expert team answers our 8 questions. Chinese Youth Daily–Freezing Point.].

[50] 萧辉. 重症科医生亲述:我们是怎样抢救危重病人的. `http://china.caixin.com/2020-02-05/101511802.html`, February 5 2020. Accessed 2020-11-26. [English translation: Xiao, Hui. Intensive care doctor describe: How do we rescue critically ill patients? Caixin.].

[51] 许冰清．1 月 6 日之后，12 天病例零新增之谜．第一财经, `https://www.yicai.com/news/100485217.html`, January 28 2020. Accessed 2020-11-28. [English translation: Xu, Bingqing. The mystery of zero new cases in 12 days after 6th of January. Yicai.].

[52] 高昱．"华南海鲜市场接触史"罗生门武汉市卫健委"双标"令人迷惑．财新网 `https://www.caixin.com/2020-02-19/101517544.html`, February 19 2020. Accessed 2020-11-28. [English translation: Gao, Yu. Rashomon about contact with the Huanan seafood market: Wuhan Municipal Health Commission's "double standard" is confusing. Caixin.].

[53] 高昱, 彭岩锋, 杨睿, 冯禹丁, and 马丹萌．独家 | 新冠病毒基因测序溯源：警报是何时拉响的．财新网 `https://archive.li/YylMt`, February 26 2020. Accessed 2020-11-27 (the original post has been removed). [English translation: Gao, Yu and Peng, Yanfeng and Yang, Rui and Feng, Yuding and Ma, Danmeng. Origin of the genetic sequencing of the novel coronavirus: When did the alarm first sound? Caixin.].

[54] 魏芙蓉, 杜萌, and 张逸凡. 武汉疫情初期, 网络直报系统为何失灵? . 新京报-剥洋葱, March 14 2020. Accessed 2020-12-01. [English translation: Wei, Furong and Du, Meng and Zhang, Yifan. Why did the direct network reporting system fail in the early stage of the Wuhan epidemic? The Beijing News.].