BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic

Qingyuan Zhao*[1], Nianqiao Ju[2], and Sergio Bacallado[1]

[1]Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge
[2]Department of Statistics, Harvard University

April 17, 2020

Abstract

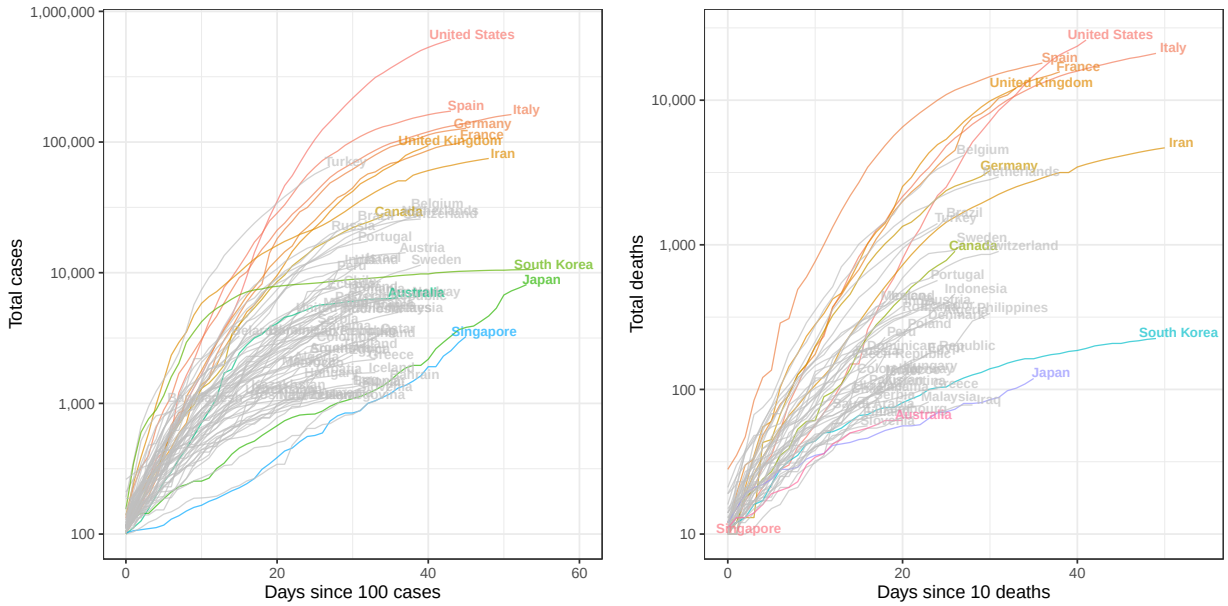
The coronavirus disease 2019 (COVID-19) has quickly grown from a regional outbreak in Wuhan, China to a global pandemic. Early estimates of the epidemic growth and incubation period of COVID-19 may have been severely biased due to sample selection. Using detailed case reports from 14 locations in and outside mainland China, we obtained 378 Wuhan-exported cases who left Wuhan before an abrupt travel quarantine. We developed a generative model we call BETS for four key epidemiological events—Beginning of exposure, End of exposure, time of Transmission, and time of Symptom onset (BETS)—and derived explicit formulas to correct for the sample selection. We gave a detailed illustration of why some early and highly influential analyses of the COVID-19 pandemic were severely biased. All our analyses, regardless of which subsample and model were being used, point to an epidemic doubling time of 2 to 2.5 days during the early outbreak in Wuhan. A Bayesian nonparametric analysis further suggests that 5% of the symptomatic cases may not develop symptoms within 14 days since infection.

1 Introduction

On December 31, 2019, the Health Commission in Wuhan, China, announced 27 cases of unknown viral pneumonia and alerted the World Health Organization. The causative pathogen was quickly identified as a novel coronavirus and the disease was later designated as the coronavirus disease 2019 (COVID-19) [4]. The regional outbreak in Wuhan quickly turned into a global pandemic. As of April 15, 2020, COVID-19 has reached almost every country in the world, infected at least 2 million people, and killed at least 130,000 [2].

Researchers around the world quickly responded to the COVID-19 outbreak. In particular, many have examined early outbreak data to estimate the initial epidemic growth, using COVID-19 cases confirmed in Wuhan or elsewhere. Two early studies published in premier medical journals by the end of January estimated that the epidemic doubling time in Wuhan was about 6 to 7 days [13, 20], but other studies appearing around the same time found that the doubling time was drastically shorter, about 2 to 3 days [16, 18, 21]. How the pandemic subsequently developed around the world seems to suggest that the latter estimates were much closer to truth. By simply plotting the cumulative cases and deaths over time, it is evident now that the number of cases (and deaths) grew

*Correspondence to: qyzhao@statslab.cam.ac.uk.



(a) Cumulative cases.

(b) Cumulative deaths.

Figure 1: Growth of the COVID-19 pandemic around the world (data retrieved from `https://www.worldometers.info/` on April 15, 2020).

more than 100 times 20 days after the first 100 cases (and 10 deaths) in countries most heavily hit by the pandemic such as Italy, Spain, and the United States (Figure 1). That growth rate almost exactly corresponds to a doubling time of 3 days. Nevertheless, to our knowledge there is no formal explanation for this drastic difference, and it might have caused confusion during the early phase of containment of COVID-19. For example, during the UK government's daily briefing on March 16, it was acknowledged that "without drastic action, cases could double every five or six days" [3]. Less than two weeks later, that number was revised to "three to four days" [1].

For infectious diseases, another key epidemiological parameter is the incubation period. Several studies have attempted to estimate the incubation period distribution of COVID-19 using cases exported from Wuhan [5, 11, 14] and the results have been instrumental in shaping guidelines to manage confirmed COVID-19 patients. For example, the interim clinical guidance for managing COVID-19 patients published by the Centers for Disease Control and Prevention (CDC) [9] quoted the results of Linton et al. [14] that "97.5% of persons with COVID-19 who develop symptoms will do so within 11.5 days of SARS-CoV-2 infection." However, as we will demonstrate below in Section 4, the design and statistical inference of these studies are highly susceptible to selection bias.

In general, there are several potential biases in early analyses of the COVID-19 pandemic:

(i) **Under-ascertainment:** Because COVID-19 is a new disease, the testing capacity was very limited during the early stage of outbreak. This may explain why Li et al. [13] under-estimated the epidemic growth as they only used cases in Wuhan who showed symptoms before January 5, 2020. Under-ascertainment also leads to under-estimation of the incubation period, as patients with longer incubation periods may be more likely to be under-ascertained during the early stages of an outbreak.

(ii) **Travel quarantine:** Wuhan is a major transportation hub in central China, but all outbound travels were abruptly halted on January 23, 2020 due to the rapid growth of the epidemic.

For studies using cases exported from Wuhan, ignoring the sample selection due to the travel quarantine leads to biased estimates of the epidemiological parameters.

(iii) **Non-random sample selection:** Because of the size and rapid growth of the COVID-19 pandemic, for most cases it is impossible to ascertain exactly when they were infected. If one simply uses the cases with known incubation period (for example, if they were known to have contact with other confirmed cases), that may create bias due to non-random sample selection. In particular, it might be less likely for cases with longer incubation period to be included in this sample, so the incubation period can be under-estimated. The abrupt travel quarantine of Wuhan created an unbiased sample, as one can observe a window of exposure for every Wuhan-exported case. However, it is still crucial to take into account under-ascertainment bias and another bias due to the epidemic growth, as explained in detail in Section 4.

In this article, we address these challenges by carefully constructing a study sample and a statistical model. We collected key epidemiological information about 1,460 confirmed COVID-19 cases across 14 locations in and outside mainland China. The health agencies in these locations have published detailed case reports since the first confirmed local case, so we do not suffer from the non-random sample selection bias described above. Section 2 describes the data collection and how we discerned the Wuhan-exported cases.

We address the sample selection due to the January 23 travel quarantine by constructing a generative statistical model. We call it the BETS model, because it models four key epidemiological events: Beginning of exposure, End of exposure, time of Transmission, and time of Symptom onset. The travel quarantine puts a constraint on the support of the observed data for Wuhan-exported cases, for which we carefully work out the selection probability and use it to adjust the likelihood function. We derive two likelihood functions, one conditional on the beginning and end of exposure and one unconditional, that can both be used to estimate the epidemic growth and the incubation period. Explicit formulae for the likelihood functions are derived under certain parametric assumptions. Detailed construction and results of the parametric model can be found in Section 3.

We then give a detailed explanation in Section 4 of why some early analyses of the COVID-19 outbreak were severely biased, including the estimation of epidemic growth by Wu et al. [20] and the estimation of incubation period by Backer et al. [5], Lauer et al. [11], Linton et al. [14]. Because these analyses did not start from a generative model, they could not correctly adjust for sample selection in the statistical inference.

In order to obtain closed-form likelihood functions in Section 3, we introduced some parametric assumptions which necessarily restrict the shape of the tail of the incubation period distribution. To avoid biased tail estimates, we model the distribution nonparametrically and also relax the other assumptions in Section 5. Because the likelihood function is no longer available in closed form, a Markov Chain Monte Carlo (MCMC) sampler is needed for Bayesian nonparametric inference. Finally, we summarize our findings and discuss potential limitations of our study in Section 6. All technical derivations can be found in the appendix; our dataset and statistical programs are publicly available as an `R` package from `https://github.com/qingyuanzhao/2019-nCov-Data`.

2 Data

2.1 Data Collection

We found 14 locations where the local health agencies have published continuous reports for every confirmed COVID-19 case since the first local case. Out of the 14 locations, 8 are cities/provinces in mainland China: Hefei, Guilin, Jinan, Shaanxi, Shenzhen, Yangzhou, Xinyang, Zhanjiang and 6 are

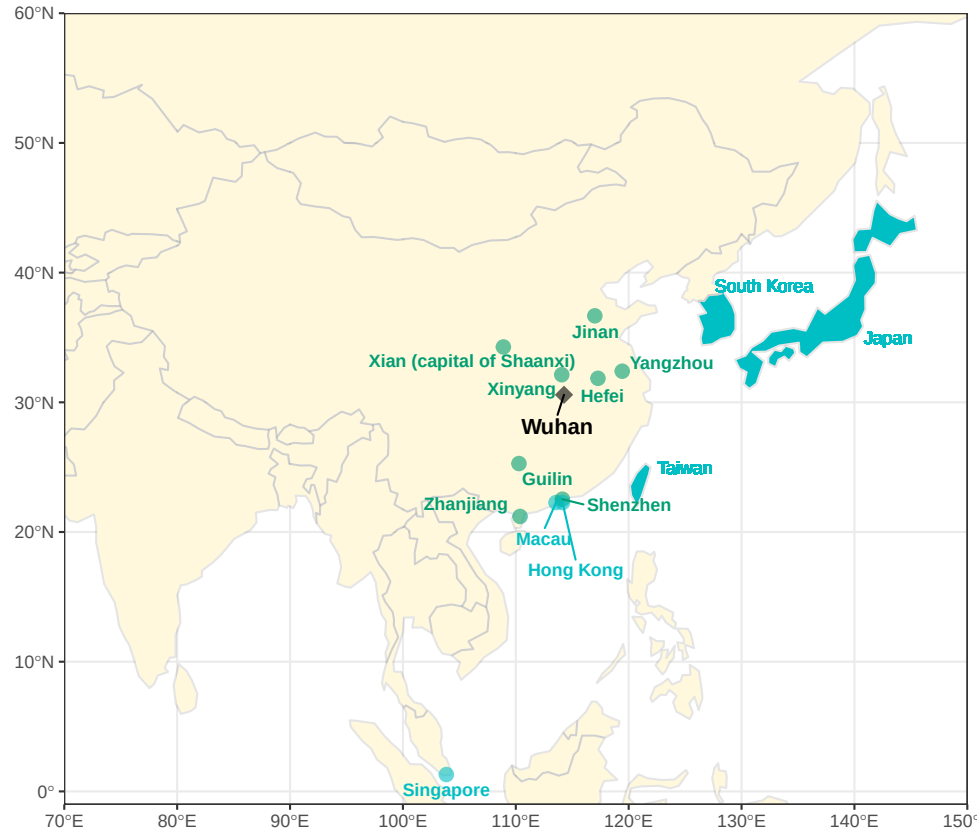


Figure 2: Geographical locations of the confirmed cases in our dataset.

countries/regions in East Asia: Hong Kong, Japan, South Korea, Macau, Singapore, and Taiwan (Figure 2). These locations have varied levels of economic development and patterns of traveling to/from Wuhan. Key information (close contact, travel history, symptom onset) of the confirmed COVID-19 cases was collected based on press releases of the official health agencies (Table 1). In total, there are 1,460 COVID-19 cases in the collected dataset.

For the mainland Chinese locations, the dataset included all the cases confirmed as of February 29, 2020. In Chinese cities outside the Hubei province, local epidemics were considered to be successfully contained by the end of February. For the international locations, the dataset included all the cases confirmed before February 15, more than three weeks after the outbound travel quarantine of Wuhan on January 23. It is thus safe to say that our dataset contains almost all Wuhan-exported cases confirmed in these locations.

2.2 Discerning Wuhan-exported cases

In total, 614 cases in our dataset had exposure to Wuhan before January 23, 2020. Because Wuhan is was the first center of epidemic outbreak and traveling from/to Wuhan was not restricted before January 23, it is reasonable to assume that the majority of these 614 cases were also infected there.

Column name	Description	Example[1]	Summary statistics
`Case`	Unique identifier for each case	HongKong-05	1460 in total
`Residence`	Nationality or residence of the case	Wuhan	21.5% reside in Wuhan
`Gender`	Gender	Male/Female	52.1%/47.7% (0.2% unknown)
`Age`	Age	63	Mean=45.6, IQR=[34, 57]
`Known Contact`	Have known epidemiological contact[2]?	Yes/No	84.7%/15.3%
`Cluster`	Relationship with other cases	Husband of HongKong-04	32.1% known
`Outside`	Transmitted outside Wuhan?[3]	Yes/Likely/No	58.5%/7.7%/33.8%
`Begin Wuhan`	Begin of stay in Wuhan (B)	30-Nov[4]	
`End Wuhan`	End of stay in Wuhan (E)	22-Jan	
`Exposure`	Period of exposure	1-Dec to 22-Jan	58.9% known period/date 8.2% known date
`Arrived`	Final arrival date at the location where confirmed a COVID-19 case	22-Jan	40.6% did not travel outside
`Symptom`	Date of symptom onset (S)	23-Jan	9.0% unknown
`Initial`	Date of first medical visit/quarantine	23-Jan	6.5% unknown
`Confirmed`	Date confirmed as a COVID-19 case	24-Jan	

Table 1: A summary of the key columns in the collected dataset.

[1]Description of this case in Hong Kong government's press release on January 24, 2020: "The other two cases are a married couple of residents of in Wuhan, a 62-year-old female [HongKong-04] and a 63-year-old male [HongKong-05], 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." (translated from Chinese).

[2]A case is considered to have known epidemiological contact if he/she had contact with people from the Hubei province or had contact with another case confirmed earlier.

[3]See the main text for the criterion we used to classify the cases.

[4]The beginning of stay is treated as November 30 if the case resides in Wuhan and has no known beginning of stay.

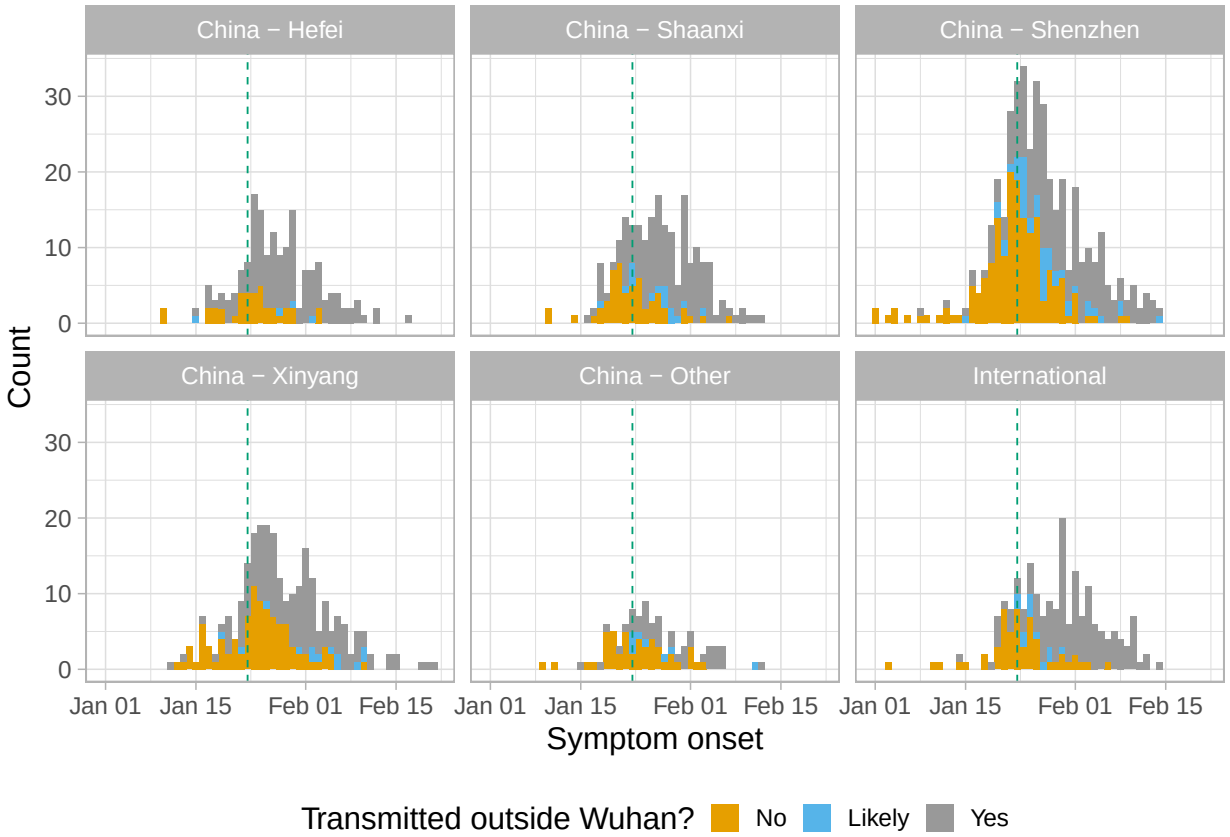


Figure 3: Epidemic curves in different locations stratified by whether the cases were transmitted outside Wuhan. "China - Other" includes four Chinese cities: Guilin, Jinan, Yangzhou, Zhanjiang; "International" includes six Asian countries/regions: Hong Kong, Japan, Korea, Macau, Singapore, Taiwan. The dashed vertical lines correspond to the abrupt travel quarantine of Wuhan from January 23, 2020.

However, some uncertainty arises if a case had contact with other confirmed cases outside their stay in Wuhan, in one of the following scenarios:

- The case already had contact with other confirmed cases before their stay in Wuhan (4 cases);
- The case had contact with other confirmed cases only after they left Wuhan but before they arrived at their destination, for example in trains or flights (4 cases);
- The case had close contact with other confirmed cases (usually family members from Wuhan) after they reached their travel destination (131 cases).

We assumed in the first two scenarios the cases were transmitted outside Wuhan. For the third scenario, it is likely that the cases were transmitted outside Wuhan, but at least one of the cases in each cluster were transmitted in Wuhan. (Two cases are considered to belong two the same cluster if they are in the family or had other recorded contact.) We used a column called `Outside` in our dataset to record our best judgment on whether the cases were transmitted outside Wuhan using the following rules:

(i) `Outside` = "Yes": Cases with no recorded stay in Wuhan between December 1, 2019 and January 23, 2020, and the 8 cases in the first two scenarios above (854 cases).

(ii) `Outside` = "Likely": Wuhan-exposed cases who did not show symptoms during the recorded stay in Wuhan and had recorded contact with another confirmed COVID-19 case with an earlier symptom onset (112 cases).

(iii) `Outside` = "No": Wuhan-exposed cases with no recorded contact with other confirmed cases, or who had the earliest symptom onset in their cluster or showed symptoms during their stay in Wuhan (494 cases).

Figure 3 shows the local epidemic curves stratified by the `Outside` column in different locations.

The dataset we collected has relatively few missing values in the key entries needed for epidemic modeling. Among the `Outside` = "No" cases, only 6.5% do not have the exact date they left Wuhan and only 8.1% have missing symptom onset date (including those showing no symptoms at the time of confirmation).

3 Statistical model and parametric inference

3.1 BETS: A generative model

We will first outline a generative model for (and named after) four key epidemiological events: the beginning of stay in Wuhan B, the end of stay in Wuhan E, the usually unobserved time of transmission T, and the time of symptom onset S (BETS). These four variables are well defined regardless of whether the person has been to Wuhan, contracted the pathogenic coronavirus, or showed symptoms of COVID-19.

Study population: Exposed to Wuhan

Consider the population of all people who stayed in Wuhan any time between 12AM December 1, 2019 (time 0) and 12AM January 24, 2020 (time L when outbound travel from Wuhan was banned, $L = 54$) in local time. We introduce the following conventions to define the population with exposure to Wuhan:

- $B = 0$: The person started their stay in Wuhan before December 2019;
- $E = \infty$: The person did not arrive in the 14 locations we are considering before the travel quarantine (time L);
- $T = \infty$: The person did not contract the pathogenic virus during their stay in Wuhan (for the purpose of this study, we need not differentiate between people who contracted the virus outside their Wuhan stay and people who never contracted the virus);
- $S = \infty$: The person did not show symptoms of COVID-19, either because they never contracted the virus or they were asymptomatic.

Because we are only considering people exposed to Wuhan, we have $B \le L$. Two other natural constraints are $B \le E$ and $T \le S$ (where we allow $\infty \le \infty$). Therefore, the support of (B, E, T, S) for the Wuhan-exposed population is

$$\mathcal{P} = \big\{(b, e, t, s) \mid b \in [0, L], e \in [b, L] \cup \{\infty\}, t \in [b, e] \cup \{\infty\}, s \in [t, \infty]\big\}. \tag{1}$$

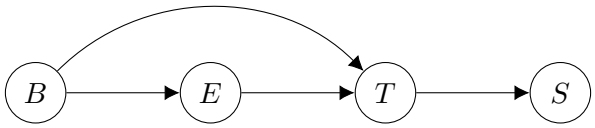


Figure 4: Directed acyclic graph (DAG) for the BETS model. B is the beginning of exposure, E is the end of exposure, T is the time of transmission, and S is the time of symptom onset.

Full data BETS model

The joint density of (B, E, T, S) can always be factorized as:

$$f(b, e, t, s) = f_B(b) \cdot f_E(e \mid b) \cdot f_T(t \mid b, e) \cdot f_S(s \mid b, e, t). \tag{2}$$

Throughout this article we will maintain two general assumptions about two conditional densities in this factorization:

Assumption 1. The conditional density $f_T(t \mid b, e)$ does not depend on b and e in the range $b \le t \le e$, so it can be written as

$$f_T(t \mid b, e) = \begin{cases} g(t), & \text{if } b < t < e, \\ 1 - \int_b^e g(x)\, dx, & \text{if } t = \infty. \end{cases} \tag{3}$$

Here $g(t) \ge 0$ models the epidemic growth in Wuhan before the citywide quarantine on January 23; it can be interpreted as the instantaneous probability of being infected in Wuhan at time t and satisfies the constraint $\int_0^L g(x)\, dx \le 1$.

Assumption 2. The conditional density $f_S(s \mid b, e, t)$ does not depend on b and e, so it can be written as

$$f_S(s \mid b, e, t) = \begin{cases} \nu \cdot h(s - t), & \text{if } s < \infty, \\ 1 - \nu, & \text{if } s = \infty. \end{cases} \tag{4}$$

Here $h(s - t)$ is the conditional density of the incubation period $S - T$ given that $S - T < \infty$ (the case is not asymptomatic), so $h(\cdot)$ satisfies $\int_0^\infty h(x)\, dx = 1$.

Assumptions 1 and 2 essentially mean that the disease transmission and progression are independent of traveling, which allows us to extend conclusions learned from the Wuhan-exported sample to the whole population. Assumption 2 may be written $S \perp\!\!\!\perp (B, E) \mid T$, which can be represented as a directed acyclic graphical (DAG) model (Figure 4) on the distribution of (B, E, T, S) [12]. Assumption 1 further restricts the dependence of T on (B, E). Under these two assumptions, the BETS model is then parameterized by two kinds of parameters: parameters for the traveling pattern $f_B(\cdot)$ and $f_E(\cdot \mid \cdot)$, and parameters for the disease transmission and progression $g(\cdot)$ and $h(\cdot)$.

Like any other assumptions in epidemic models, Assumptions 1 and 2 represent approximations to the underlying dynamics. Assumptions 1 and 2 can be violated if, for example, short-term visitors were exposed to more infectious cases or if people were less likely to travel if they felt sick. Nevertheless, we think they are reasonable approximations to the reality during the initial outbreak, when little was known about the new infectious disease.

Parametric assumptions

Assumptions 1 and 2 are general assumptions on the dependence of T and S on B and E. We consider two parametric assumptions that simplify the interpretation of our results:

Assumption 3. The probability of contracting the virus in Wuhan was increasing exponentially before the quarantine:

$$g(t) = g_{\kappa,r}(t) \stackrel{\Delta}{=} \kappa \cdot \exp(rt),\ t \leq L, \tag{5}$$

where (κ, r) satisfies $\int_0^L g_{\kappa,r}(t)\, dt \leq 1$.

Assumption 4. The incubation period $T - S$, given that it is finite (the case is not asymptomatic), follows a Gamma distribution with shape $\alpha > 0$ and rate $\beta > 0$:

$$h(s-t) = h_{\alpha,\beta}(s-t) \stackrel{\Delta}{=} \frac{\beta^\alpha}{\Gamma(\alpha)}(s-t)^{\alpha-1} \exp\{-\beta(s-t)\}. \tag{6}$$

Assumption 3 says that the epidemic size in Wuhan was growing exponentially before the quarantine, which is a common assumption for early epidemic outbreaks. We think it is quite reasonable given that little was known about the novel coronavirus before January 23. Assumption 4 restricts the density function $h(\cdot)$ to the Gamma family, which is commonly used to model the distribution of the incubation period. These two assumptions will be used later in this and the next sections to calculate closed-form likelihood functions. Later in Section 5, we will relax the parametric assumptions to allow more flexible patterns for the epidemic growth and more general distributions of the incubation period.

3.2 Accounting for sample selection in the likelihood

Study sample: Wuhan-exported cases

To use Wuhan-exported cases to study the epidemic growth and incubation period, it is crucial to consider the effect of sample selection on Wuhan-exported cases. Using the notation above, the Wuhan-exported cases confirmed in the 14 locations we consider can be written as an event $(B, E, T, S) \in \mathcal{D}$ where

$$\mathcal{D} = \{(b, e, t, s) \in \mathcal{P} \mid b \leq t \leq e \leq L, t \leq s < \infty\}. \tag{7}$$

Compared to the full population $\mathcal{P}$ of people with exposure to Wuhan in (1), this set makes three further restrictions:

(i) $B \leq T \leq E$, because we only use cases who contracted the virus during their stay in Wuhan;

(ii) $E \leq L$, because the case can only be observed in the dataset if he/she left Wuhan before the travel quarantine;

(iii) $S < \infty$, because not all locations report asymptomatic cases, which motivates us to only consider COVID-19 cases who have shown symptoms.

Selection-adjusted likelihood functions

In an ideal world where we could take independent observations from the exposed population $\mathcal{P}$, the likelihood function would be given by a product of the density $f(B_i, E_i, T_i, S_i)$ in (2) over i in the sample. However, that is not the case for the initial COVID-19 outbreak in Wuhan. Because of limited testing capacity in the beginning of the outbreak, many COVID-19 patients were not identified.

Instead, we have obtained a high-quality dataset of Wuhan-exported cases which can be considered as "shadows" of the outbreak in Wuhan. In order to use this dataset, it is crucial that the statistical inference takes into account the sample selection (because we do not have independent observations from $\mathcal{P}$). In other words, statistical inference should be based on the conditional density:

$$f(b, e, t, s \mid \mathcal{D}) \triangleq f\big(b, e, t, s \mid (B, E, T, S) \in \mathcal{D}\big) = \frac{f(b, e, t, s) \cdot 1_{\{(b,e,t,s)\in\mathcal{D}\}}}{\mathbb{P}\left((B, E, T, S) \in \mathcal{D}\right)}, \tag{8}$$

where $1_{\{\cdot\}}$ is the indicator function. Once this conditional density is obtained, we can then use the product

$$\prod_{i=1}^{n} f\big(B_i, E_i, T_i, S_i \mid \mathcal{D}\big), \tag{9}$$

as the likelihood function for frequentist or Bayesian inference, under the assumption that we have observed an independent and identically distributed sample $(B_i, E_i, T_i, S_i),\ i = 1, \ldots, n$ from the conditional density.

However, in our dataset the time of transmission T is usually not observed. We can either treat T as a latent variable and maximize the likelihood over both the modeling parameters and the unobserved T_i, or simply marginalize over T in the full data likelihood and use the following observed data likelihood,

$$L_{\text{uncond}}(\theta) = \prod_{i=1}^{n} \int_{(B_i,E_i,t,S_i)\in\mathcal{D}} f\big(B_i, E_i, t, S_i \mid \mathcal{D}\big)\, dt, \tag{10}$$

where $\theta = (f_B(\cdot), f_E(\cdot|\cdot), g(\cdot), h(\cdot))$ contains all the parameters of interest.

We can also condition on (B, E) to formulate a conditional likelihood function that does not depend on the marginal distribution of (B, E):

$$L_{\text{cond}}(\theta) = \prod_{i=1}^{n} \int_{(B_i,E_i,t,S_i)\in\mathcal{D}} f_{T,S}\big(t, S_i \mid B_i, E_i, \mathcal{D}\big)\, dt, \tag{11}$$

where $\theta = (g(\cdot), h(\cdot))$ and

$$f_{T,S}(t, s \mid b, e, \mathcal{D}) \triangleq f_{T,S}(t, s \mid b, e, (B, E, T, S) \in \mathcal{D}) = \frac{f_{T,S}(t, s \mid b, e) \cdot 1_{\{(b,e,t,s)\in\mathcal{D}}}{\mathbb{P}((B, E, T, S) \in \mathcal{D} \mid B = b, E = e)}. \tag{12}$$

The information about the epidemic growth $g(\cdot)$ and the incubation period $h(\cdot)$ contained in the density $f_{B,E}(B, E \mid \mathcal{D})$ is not used in this likelihood, but the benefit is that the conditional likelihood does not require us to specify the traveling pattern $f_B(\cdot)$ and $f_E(\cdot \mid \cdot)$.

Computing the selection probability

Next we derive the likelihood functions (10) and (11) under additional parametric modeling assumptions on the traveling pattern $f_B(\cdot)$ and $f_E(\cdot \mid \cdot)$. The first technical problem here is to compute the denominators in (8) and (12). This is straightforward for the conditional likelihood:

Lemma 1. *Under Assumptions 1 and 2, for* $(b, e, t, s) \in \mathcal{D}$,

$$\mathbb{P}((B, E, T, S) \in \mathcal{D} \mid B = b, E = e) = \nu[G(e) - G(b)], \text{ and } f_{T,S}(t, s \mid b, e, \mathcal{D}) = \frac{g(t)h(s-t)}{G(e) - G(b)}. \quad (13)$$

where $G(t) = \int_{-\infty}^{t} g(x)\,dx$. *If we additionally assume* $g(t)$ *is growing exponentially (Assumption 3), we have*

$$f_{T,S}(t, s \mid b, e, \mathcal{D}) = \begin{cases} \dfrac{r\exp(rt)}{\exp(re) - \exp(rb)}\, h(s-t), & \text{for } r \neq 0, \\ \dfrac{\exp(rt)}{e-b}\, h(s-t), & \text{for } r = 0. \end{cases} \quad (14)$$

An important observation is that (14) does not depend on ν (proportion of symptomatic cases) and κ (absolute scale of the epidemic).

For the denominator in the unconditional likelihood we need to integrate $\mathbb{P}((B, E, T, S) \in \mathcal{D} \mid B = b, E = e)$ over the marginal distribution of B and E. We cannot further simplify the integral without making assumptions on $f_B(b)$ and $f_E(e \mid b)$. For this purpose we make the following two assumptions which heuristically say that the travel pattern is stable during the study period:

Assumption 5. The beginning of stay in Wuhan B, conditioning on $0 \leq B \leq L$, follows a uniform distribution from 0 to L. More specifically,

$$f_B(b) = \begin{cases} 1 - \pi, & \text{for } b = 0, \\ \pi/L, & \text{for } 0 < b \leq L, \end{cases} \quad (15)$$

where $0 \leq \pi \leq 1$ is the proportion of visitors (non-residents of Wuhan) in the Wuhan-exposed population.

Assumption 6. The end of stay E follows an uniform distribution from B to L given $E \leq L$, with rate depending on whether the person resides in Wuhan:

$$f_E(e \mid b = 0) = \begin{cases} \lambda_W, & \text{if } 0 \leq e \leq L, \\ 1 - L\lambda_W, & \text{if } e = \infty, \end{cases}, \quad f_E(e \mid b, b > 0) = \begin{cases} \lambda_V, & \text{if } b \leq e \leq L, \\ 1 - (L-b)\lambda_V, & \text{if } e = \infty, \end{cases} \quad (16)$$

where the parameters $\lambda_W, \lambda_V \leq 1/L$.

For $b > 0$, this assumption implies that $\mathbb{P}(E = \infty | b, b > 0) = b\lambda_V + (1 - L\lambda_V)$ increases as b increases. This is consistent with our intuition that the later someone arrives in Wuhan, the more likely the person stays there after the travel quarantine on January 23.

By using the parametric forms (5), (15), (16) when integrating $\mathbb{P}((B, E, T, S) \in \mathcal{D} \mid B = b, E = e)$ and using the approximation $(1 + rL)/\exp(rL) \approx 0$ for $rL > 5$, after some algebra we obtain

Lemma 2. *Under Assumptions 1 to 3, 5 and 6, for* $r > 5/L$, *the selection probability is given by*

$$\mathbb{P}((B, E, T, S) \in \mathcal{D}) \approx \frac{\kappa \exp(rL)\nu}{r^2}\Big[(1-\pi)\lambda_W + \pi\lambda_V\Big(1 - \frac{2}{rL}\Big)\Big],$$

and for $(b, e, t, s) \in \mathcal{D}$, *the conditional density is given by*

$$f(b, e, t, s \mid \mathcal{D}) \approx r^2 \cdot \frac{[1_{\{b=0\}} + (\rho/L)1_{\{b>0\}}] \cdot \exp(rt)}{\Big[1 + \rho(1 - 2/(rL))\Big] \cdot \exp(rL)} \cdot h(s-t), \quad (17)$$

where $\rho = (\lambda_V/\lambda_W) \cdot \pi/(1-\pi)$.

Similar to (14), the conditional density (17) does not depend on ν and κ. Moreover, it only depends on the traveling parameters π, λ_V and λ_W through a single transformed parameter ρ. The approximation $(1+rL)/\exp(rL) \approx 0$ we used to obtain the analytical formulae in Lemma 2 is quite reasonable for $rL > 5$ (if the doubling time is 4 days, $rL = \log(2)/4 \times 54 = 9.34$).

Observed data likelihood

As explained after equation (9), we cannot immediately use the conditional density (14) or (17) for statistical inference because we do not observe the time of transmission T. The final step in the derivation of our likelihood function is to marginalize over t in the density functions. The parametric form of $h(\cdot)$ in Assumption 4 allows us to derive closed-form formulae.

Proposition 1. *Under Assumptions 1 to 4, the observed data conditional likelihood* (11) *is given by*

$$L_{cond}(r,\alpha,\beta) = \begin{cases} r^n\Big(\dfrac{\beta}{\beta+r}\Big)^{n\alpha} \cdot \displaystyle\prod_{i=1}^{n} \frac{\exp(rS_i)\big[H_{\alpha,\beta+r}(S_i-B_i) - H_{\alpha,\beta+r}((S_i-E_i)_+)\big]}{\exp(rE_i)-\exp(rB_i)}, & \text{for } r>0, \\ \displaystyle\prod_{i=1}^{n} \frac{H_{\alpha,\beta}(S_i-B_i) - H_{\alpha,\beta}((S_i-E_i)_+)}{E_i-B_i}, & \text{for } r=0, \end{cases} \tag{18}$$

where $H_{\alpha,\beta}(\cdot)$ is the cumulative distribution function of the Gamma distribution with shape α and rate β and $(x)_+ = \max(x,0)$ is the positive part of x. Under Assumptions 1 to 6, the observed data unconditional likelihood (10) *for $r > 5/L$ is approximately given by*

$$\begin{aligned} L_{uncond}(\rho,r,\alpha,\beta) \approx r^{2n}\Big(\frac{\beta}{\beta+r}\Big)^{n\alpha} \cdot \prod_{i=1}^{n} \Bigg\{ & \frac{1_{\{B_i=0\}} + (\rho/L)1_{\{B_i>0\}}}{1+\rho(1-2/(rL))} \exp\big\{r(S_i-L)\big\} \\ & \times \big[H_{\alpha,\beta+r}(S_i-B_i) - H_{\alpha,\beta+r}((S_i-E_i)_+)\big]\Bigg\}. \end{aligned} \tag{19}$$

It is worthwhile to point out that if $r = 0$ (the epidemic was stationary), our conditional likelihood function $L_{\text{cond}}(r,\alpha,\beta)$ reduces to the likelihood function for interval-censored exposure in Reich et al. [17]. However, COVID-19 was growing quickly during its early outbreak in Wuhan, so the growth exponent r is very different from 0. It is thus inappropriate to use the likelihood $L_{\text{cond}}(0,\alpha,\beta)$ to estimate the incubation period of COVID-19, as done in some previous analyses also using Wuhan-exported cases [5, 11, 14]. See Section 4.2 for further discussion and an illustration of the bias due to ignoring the epidemic growth.

3.3 Results of the parametric inference

Implementation

To fit the statistical model, we used the 378 cases in our dataset that satisfy our sample selection criterion in Section 3.2 and do not have missing symptom onset. We fitted separate models for different locations and compare the results across the locations.

As the model in Section 3 is a regular parametric model, we performed the usual frequentist inference using the likelihood function (19). In particular, point estimators of the parameters (ρ,r,α,β) were obtained by maximizing the likelihood function (19), and confidence intervals for the parameters were obtained by inverting the likelihood ratio χ^2-test. As we are more interested in quantiles of the incubation period instead of the shape and rate parameters, we parametrized the

Location	Sample size	ρ	Doubling time (in days)	Incubation period	
				Median	95% quantile
Conditional likelihood					
China - Hefei	34	Not estimated	2.1 (1.2–3.7)	4.3 (2.9–6.0)	12.0 (9.1–17.3)
China - Shaanxi	53	Not estimated	1.7 (1.0–2.8)	4.5 (3.1–6.2)	14.6 (11.5–19.8)
China - Shenzhen	129	Not estimated	2.2 (1.7–3.0)	3.5 (2.8–4.3)	11.2 (9.5–13.6)
China - Xinyang	74	Not estimated	2.3 (1.5–3.5)	6.8 (5.4–8.2)	16.4 (13.8–20.1)
China - Other	42	Not estimated	2.0 (1.1–3.4)	5.1 (3.6–6.7)	12.3 (9.8–16.4)
International	46	Not estimated	2.1 (1.4–3.4)	3.8 (2.5–5.3)	10.9 (8.4–15.1)
All locations	378	Not estimated	2.1 (1.8–2.5)	4.5 (4.0–5.0)	13.4 (12.2–14.8)
All except Xinyang	304	Not estimated	2.1 (1.7–2.5)	4.0 (3.5–4.6)	12.2 (11.0–13.7)
Unconditional likelihood					
China - Hefei	34	0.40 (0.18–0.82)	1.8 (1.4–2.4)	4.1 (2.8– 5.5)	11.9 (9.0–17.2)
China - Shaanxi	53	0.24 (0.11–0.46)	2.5 (2.0–3.1)	5.3 (3.9– 6.8)	15.0 (12.0–20.0)
China - Shenzhen	129	0.75 (0.52–1.06)	2.4 (2.1–2.8)	3.6 (2.9– 4.3)	11.3 (9.6–13.7)
China - Xinyang	74	0.45 (0.27–0.74)	2.4 (2.0–2.9)	6.8 (5.6– 8.1)	16.4 (13.9–20.2)
China - Other	42	0.45 (0.22–0.86)	2.1 (1.7–2.8)	5.3 (4.0– 6.6)	12.4 (10.0–16.4)
International	46	0.14 (0.05–0.32)	2.0 (1.6–2.6)	3.7 (2.5– 5.0)	10.8 (8.4–15.1)
All locations	378	0.45 (0.36–0.56)	2.3 (2.1–2.5)	4.6 (4.1– 5.1)	13.5 (12.3–14.9)
All except Xinyang	304	0.45 (0.35–0.57)	2.2 (2.1–2.5)	4.1 (3.7– 4.6)	12.3 (11.1–13.8)

Table 2: Results of the parametric inference. For each location and parameter, the maximum likelihood estimator and the 95% confidence interval (in brackets) based on inverting the likelihood ratio test are reported.

Gamma distribution in Assumption 4 by its median and 95% quantile and mapped them to α and β when calculating the likelihood function. The growth exponent r was also transformed to the more interpretable doubling time (in days) using doubling time $= \log(2)/r$.

Because we only observed the date instead of the exact time for B, E, and S, we applied a simple transformation before computing the likelihood function. Instead of using the integer date which corresponds to the end of a day, we used $B - 3/4$, $E - 1/4$, and $S - 1/2$ in places of B, E, and S to compute (18) and (19). This transformation also avoids a singularity in the likelihood function when B and E are exactly equal.

Results

Results of the parametric model in Section 3 are reported in Table 2. We give some remarks about the results:

(i) There is considerable heterogeneity of the estimated ρ (a parameter capturing the traveling pattern) using the unconditional likelihood. This is not surprising given that the locations we are considering are different in many ways.

(ii) Regardless of the location, our model shows that the epidemic doubling time in Wuhan was less than 3 days. There is no substantial heterogeneity among estimates in different locations.

(iii) The estimated incubation periods are similar for most locations except Xinyang, a less developed city neighboring the Hubei province.

(iv) The conditional likelihood (18) and unconditional likelihood (19) give very similar results. Confidence intervals computed using the unconditional likelihood are slightly shorter than those computed using the conditional likelihood.

In conclusion, inferences based on our parametric model suggest that the initial doubling time of the COVID-19 epidemic in Wuhan was between 2 to 2.5 days, the median incubation period of COVID-19 is around 4 days, and the 95% quantile of the incubation period is between 11 to 15 days.

4 Why some previous COVID-19 analyses were severely biased

4.1 Estimating the epidemic growth: Bias due to sample selection

Like the present study, a highly influential article published in the *Lancet* in late January also used Wuhan-exported cases to estimate the epidemic growth during the early outbreak [20]. However, their estimated doubling time was 6.4 days (95% credible interval: 5.8–7.1), drastically higher than the estimates in Table 2.

A closer look at the model in Wu et al. [20] shows that the most likely reason is that their model did not consider how sample selection (in particular, the travel quarantine of Wuhan) changes the likelihood function. This issue is best illustrated by examining the marginal distribution of symptom onset in Wuhan-exported cases, which can be obtained by integrating the conditional density (17) obtained earlier:

Proposition 2. *Under Assumptions 1 to 3, 5 and 6, the marginal density of T given $(B, E, T, S) \in \mathcal{D}$ for $r > 5/L$ is approximately given by*

$$f_T(t \mid \mathcal{D}) \overset{\propto}{\sim} (L - t)\exp(rt) \cdot 1_{\{t \le L\}}, \tag{20}$$

where $\overset{\propto}{\sim}$ means approximately proportional to. If in addition the incubation period $S - T$ follows a Gamma(α, β) distribution (Assumption 4), the marginal density of S of the exported cases is given by

$$f_S(s \mid \mathcal{D}) \overset{\propto}{\sim} \exp(rs) \cdot \Big\{(L - s)[1 - H_{\alpha,\beta+r}((s - L)_+)] + \frac{\alpha}{\beta + r}[1 - H_{\alpha+1,\beta+r}((s - L)_+)]\Big\}, \tag{21}$$

and as a consequence,

$$f_S(s \mid \mathcal{D}) \overset{\propto}{\sim} \exp(rs) \cdot \Big(L + \frac{\alpha}{\beta + r} - s\Big) \text{ for } s \le L. \tag{22}$$

Figure 5 shows the histogram of the symptom onset of the Wuhan-exported cases in our dataset and the theoretical fit based on (21) and the maximum likelihood estimator in Table 2 using all the locations ($r = 0.30, \alpha = 1.86, \beta = 0.33$). The theoretical density provided good fit to the observed distribution of S (Pearson's χ^2 goodness-of-fit test: p-value = 0.94).

Wu et al. [20] fitted a Susceptible-Exposed-Infectious-Recovered (SEIR) model using Wuhan-exported cases but did not consider sample selection due to the travel quarantine. In the early phase of epidemic outbreaks, the SEIR model can be well approximated by an exponential growth for cases in Wuhan:

$$f_S(s) \propto \exp(rs).$$

However, Proposition 2, in particular equation (22), shows that the marginal distribution of S in Wuhan-exported cases $f_S(s \mid \mathcal{D})$ does not follow the same exponential growth as $f_S(s)$.

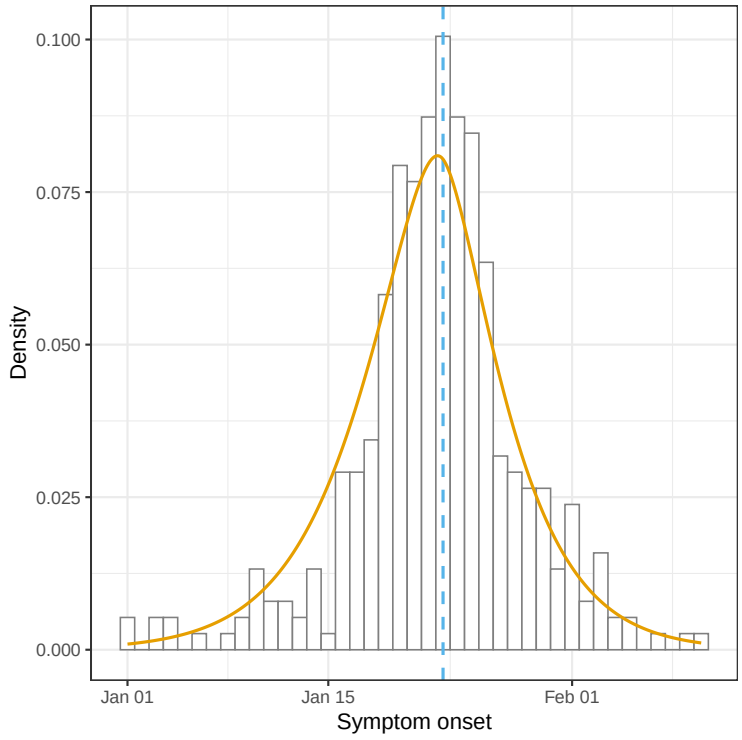


Figure 5: Marginal distribution of symptom onset of Wuhan-exported COVID-19 cases. Histogram: Density of the symptom onset date of the Wuhan-exported cases in our dataset; Orange curve: Theoretical fit based on (21); Blue dashed line: Date of travel quarantine for Wuhan (January 23, 2020).

Equation (22) not only shows that fitting a simple exponential growth to the initial symptom onsets among Wuhan-exported cases will under-estimate the epidemic growth r, we can also use it to derive a simple bias-correction formula. By using the first-order Taylor expansion,

$$\begin{aligned}\log f_S(s \mid \mathcal{D}) &\approx rs + \log\left(L + \frac{\alpha}{\beta + r} - s\right) + \text{constant} \\ &\approx rs + \log\left(\frac{\alpha}{\beta + r}\right) + \frac{L - s}{\alpha/(\beta + r)} + \text{constant} \\ &= \left[r - \frac{\beta + r}{\alpha}\right] s + \text{constant}.\end{aligned}$$

Therefore the under-estimation bias is about $(\beta + r)/\alpha \geq \beta/\alpha = 1/\mathbb{E}[S - T]$. This means that if the mean incubation period is 5 days, fitting a simple exponential growth could under-estimate r by as much as 0.2!

Using Wuhan-exported cases confirmed outside Mainland China by January 28, 2020, Wu et al. [20] estimated that the doubling time of COVID-19 was about 6.4 days, which corresponds to $r = \log(2)/6.2 \approx 0.11$. With the above correction, the actual r should be at least $0.11 + 0.2 \approx 0.31$, or doubling time of 2.2 days. This is very close to our estimates in Table 2. Although the calculations here are inexact, they clearly demonstrate that ignoring the sample selection due to travel quarantine can lead to substantial under-estimation of the epidemic growth.

4.2 Estimating the incubation period: When two biases do not "balance out"

Like the present study, several influential articles also estimated the incubation period of COVID-19 using Wuhan-exported cases [5, 11, 14]. Their results are roughly in line with our estimates in Table 2, but a closer look shows that the existing methods actually suffer from two biases:

(i) **Under-ascertainment bias:** The three previous studies only used Wuhan-exported cases confirmed before the end of January. In our dataset, about 70% of the Wuhan-exported cases were confirmed by that time. However, the other 30% would have an incubation period of at least 8 days as they must have left Wuhan before January 23. The lack of ascertainment leads to under-estimation of the incubation period.

(ii) **Epidemic growth bias:** The three previous studies all used the interval-censored likelihood function for the incubation period in Reich et al. [17]. As discussed after Proposition 1, this likelihood corresponds to our conditional likelihood $L_{\text{cond}}(\alpha, \beta)$ with r fixed at 0 and thus does not account for the rapid growth of COVID-19. Intuitively, a person in Wuhan has much higher prior probability of contracting the virus on January 20 than on January 1, but the likelihood function in Reich et al. [17] does not take that into account. Ignoring the epidemic growth leads to over-estimation of the incubation period.

It is possible to correct for the under-ascertainment by further conditioning on $S \leq M$ (M is some truncation time) in our likelihood function.

Proposition 3. *Under Assumptions 1 and 2, for $(b, e, t, s) \in \mathcal{D}$ and $s \leq M$,*

$$f_{T,S}(t, s \mid b, e, \mathcal{D}, S \leq M) = \frac{g(t)h(s-t)}{\int_b^{\max(e,s)} g(t)H(M-t)\,dt}, \tag{23}$$

where $H(s) = \int_0^s h(x)\,dx$ is the distribution function of the incubation period. Furthermore, under the exponential growth model (Assumption 3) and Gamma-distributed incubation period (Assumption 4), the conditional observed data likelihood under the right truncation $S \leq M$ is given by

$$\begin{aligned} &L_{cond,trunc}(r, \alpha, \beta; M) \\ &= \begin{cases} r^n \Big(\frac{\beta}{\beta + r}\Big)^{n\alpha} \prod_{i=1}^{n} \frac{\exp\{r(S_i - M)\}[H_{\alpha,\beta+r}(S_i - B_i) - H_{\alpha,\beta+r}((S_i - E_i)_+)]}{Z_r(M - B_i) - Z_r((M - E_i)_+)}, & \text{if } r \neq 0, \\ \prod_{i=1}^{n} \frac{H_{\alpha,\beta}(S_i - B_i) - H_{\alpha,\beta}((S_i - E_i)_+)}{Z_0(M - B_i) - Z_0((M - E_i)_+)}, & \text{for } r = 0, \end{cases} \end{aligned} \tag{24}$$

where

$$Z_r(x) = \begin{cases} \Big(\frac{\beta}{\beta + r}\Big)^{\alpha} H_{\alpha,\beta+r}(x) - \exp(-rx) H_{\alpha,\beta}(x), & \text{for } r \neq 0, \\ x H_{\alpha,\beta}(x) - \Big(\frac{\alpha}{\beta}\Big) H_{\alpha+1,\beta}(x), & \text{for } r = 0. \end{cases}$$

It is straightforward to show that $L_{\text{cond,trunc}}(r, \alpha, \beta; M)$ reduces to the conditional likelihood $L_{\text{cond}}(r, \alpha, \beta)$ without the right truncation in (18) when $M \to \infty$.

We demonstrate the two kinds of biases in the estimation of the incubation period using a retrospective experiment. In this experiment, we assumed the incubation period follows a Gamma distribution and estimated its median and the 95% quantile by maximizing one of the following three likelihood functions:

(i) **Adjusting for nothing:** This is the likelihood function in Reich et al. [17] that is equal to our $L_{\text{cond}}(0, \alpha, \beta)$ that sets $r = 0$.

(ii) **Adjusting for growth:** This is our conditional likelihood function $L_{\text{cond}}(r, \alpha, \beta)$.

(iii) **Adjusting for both growth and ascertainment:** This is our conditional likelihood $L_{\text{cond,trunc}}(r, \alpha, \beta; M)$ with adjustment for sample selection due to the right truncation $S \leq M$.

For each day from January 23 to February 18, we estimated the incubation distribution using Wuhan-exported cases in our dataset confirmed by that day. For the third method, we choose M to be a week prior to the truncation date for confirmation, as most Wuhan-exported cases were confirmed within a week of symptom onset. Figure 6 shows the estimated medians and 95% quantiles of the incubation period of COVID-19, with pointwise confidence intervals in the plot computed using the basic nonparametric bootstrap with 1000 resamples [8].

The under-ascertainment bias can be clearly visualized from the dotted blue curves in Figure 6. Had we fitted our conditional likelihood function $L_{\text{cond}}(r, \alpha, \beta)$ using cases confirmed by January 31 (265 cases), the estimated median incubation period would be 3.5 days and the 95% quantile would be 9.5 days. In comparison, when the entire dataset is used, the estimated median and 95% quantile are 4.6 days and 13.5 days (Table 2).

The over-estimation due to ignoring the epidemic growth is even more dramatic. Had we fitted the incubation period using the likelihood function in Reich et al. [17] (the same as setting $r = 0$ in our conditional likelihood) to all the cases in our dataset (387 cases), the estimated median incubation period would be 9.2 days and the 95% quantile would be a whopping 24.9 days!

The truncation-corrected conditional likelihood $L_{\text{cond,trunc}}(r, \alpha, \beta; M)$ derived in Proposition 3 successfully corrected for the under-ascertainment bias. The estimated median and 95% quantile of the incubation using $L_{\text{cond,trunc}}(r, \alpha, \beta; M)$ were roughly unbiased starting from the end of January. Had we fitted this likelihood using all cases confirmed by January 31 and having shown symptoms a week prior (220 cases), the estimated median incubation period would be 4.8 days (95% CI: 3.0 to 6.0) and the estimated 95% quantile would be 14.4 days (95% CI: 6.7 to 18.5). These estimates are less precise than the estimates obtained using the entire dataset (Table 2), but they correctly reflect the uncertainty due to the under-ascertainment. In contrast, using the wrong likelihood functions not only results in biased point estimates but also narrow and misleading confidence intervals.

Because the under-ascertainment bias and epidemic growth bias are towards opposite directions, coincidentally they were almost "balanced out" in the previous studies. As a consequence, their estimates were not drastically different from ours. To be fair in this criticism, these previous studies all acknowledged that under-ascertainment of mild cases could bias their analyses. Backer et al. [5] mentioned the over-estimation due to ignoring epidemic growth in their discussion. Linton et al. [14] attempted to use a formula to correct for under-ascertainment which bears some similarity to (23), which resulted in slightly longer estimates of the incubation period. However, they did not give any justification to the formula and we could not derive it from our generative model. Nevertheless, our experiments in Figure 6 clearly show that these early estimates of the incubation period (especially their tail estimates) are unreliable for making policy decisions.

5 Nonparametric inference

5.1 Time discretization

So far we have used parametric assumptions (e.g. Gamma-distributed incubation period) to explicitly derive likelihood functions for the observed data. To assess the robustness of our results, next we

Figure 6: An illustration of two kinds of biases in the estimation of the incubation period of COVID-19. The curves (region) in the plot are maximum likelihood estimators (and bootstrap confidence intervals) using three likelihood functions and cases confirmed by each day. Likelihood functions used in this experiment are: $L_{\text{cond}}(0, \alpha, \beta)$ (dashed orange), $L_{\text{cond}}(r, \alpha, \beta)$ (dotted blue), and $L_{\text{cond,trunc}}(r, \alpha, \beta; M)$ (solid green). Results of some previous studies [5, 11, 14] using the conditional likelihood with r set to 0 are also shown in the Figure.
(Lauer et al. [11] did not report an estimated 95% quantile of the incubation period. Here we imputed it based on the reported median and 97.5% quantile, assuming a Gamma distribution for the incubation period. Although Lauer et al. [11] used COVID-19 cases confirmed as late as late February, only 4 out of their 181 cases were confirmed in February. In this Figure it is thus treated as using cases confirmed up till February 1.)

will relax some of these parametric assumptions. In particular, we will model the distribution of the incubation period nonparametrically so the tail probabilities are not determined by any parametric form. Because analytic forms of the sample selection probabilities $\mathbb{P}((b,e,t,s) \in \mathcal{D} \mid b,e)$ and $\mathbb{P}((b,e,t,s) \in \mathcal{D})$ are generally unavailable, we will put prior distributions on the model parameters and use Markov Chain Monte Carlo (MCMC) to compute their posterior distributions.

We start by discretizing all the time variables in the model, which are measured in days. This will simplify the Bayesian computation. Instead of working with continuous time $(B,E,T,S) \in \mathcal{P}$, we use the discretization:

$$B^* = \lceil B \rceil, E^* = \lceil E \rceil, T^* = \lceil T \rceil, S^* = \lceil S \rceil,$$

where $\lceil \cdot \rceil$ is the ceiling function ($\lceil x \rceil$ is the smallest integer larger than x). The support of (B^*, E^*, T^*, S^*) is then $\mathcal{P}$, the set of all 4-tuples of integers and ∞. The general continuous distributions in Assumptions 1 and 2 can be modified accordingly:

$$\mathbb{P}(T^* = t^* \mid B^* = b^*, E^* = e^*) = \begin{cases} g^*(t^*), & \text{if } b^* \le t^* \le e^*, \\ 1 - \sum_{t^*=b^*}^{e^*} g^*(t^*), & \text{if } t^* = \infty; \end{cases}$$

$$\mathbb{P}(S^* = s^* \mid B^* = b^*, E^* = e^*, T^* = t^*) = \begin{cases} \nu \cdot h^*(s^* - t^*), & \text{if } t^* \le s^* < \infty, \\ 1 - \nu, & \text{if } s^* = \infty, \end{cases}$$

where $g^*(\cdot)$ satisfies $\sum_{x^*=0}^{L} g^*(x^*) \le 1$ and $h^*(\cdot)$ is a probability mass function on nonnegative integers: $\sum_{x^*=0}^{\infty} h^*(x^*) = 1$.

5.2 Relaxing the parametric assumptions

Our parametric assumptions (Assumptions 3 to 6) on the distribution of (B,E,T,S) can be translated to the following assumptions on (B^*,E^*,T^*,S^*) after discretization:

$$g^*(t^*) \approx g_{\kappa,r}(t^*) = \kappa \exp(r t^*), \; h^*(t^* - s^*) \approx h_{\alpha,\beta}(t^* - s^*),$$

$$\mathbb{P}(B^* = b^*) = \begin{cases} (1-\pi), & \text{for } b^* = 0, \\ \pi/L, & \text{for } b^* = 1, \ldots, L, \end{cases}$$

and

$$\mathbb{P}(E^* = e^* \mid B^* = b^*) = \begin{cases} \lambda_{b^*}, & \text{for } b^* \le e^* \le L, \\ 1 - (L - b^* + 1)\lambda_{b^*}, & \text{for } e^* = \infty. \end{cases}$$

where $\lambda_0 = \lambda_W$ and $\lambda_1 = \cdots = \lambda_L = \lambda_V$.

In the nonparametric model we consider the following relaxations:

(i) **Nonparametric distribution for the incubation period:** Besides putting a prior encouraging smoothness and log-concavity, we do not put any parametric restrictions on the distribution of the incubation period.

(ii) **Two-stage exponential growth:** Human-to-human transmissibility of COVID-19 is first confirmed to the public in the evening of January 20. We modify the exponential growth model to allow for a different growth exponent after January 20:

$$g^*(t^*) = g^*_{\kappa,r_1,r_2}(t^*) = \begin{cases} \kappa \exp(r_1 t^*) & \text{if } t \le L_1, \\ \kappa \exp(r_2(t^* - L_1) + r_1^* L_1) & \text{if } L_1 < t \le L_2, \end{cases}$$

where $L_1 = 51$ (January 20) and $L_2 = L = 54$ (January 23). The simple exponential growth model is a special case of this model with both L_1 and L_2 set to L.

(iii) **Geometric distribution for $E^* \mid B^*$:** As a sensitivity analysis to our assumption that $E^* \mid B^*$ is uniformly distributed between B^* and L, this relaxation assumes a geometric distribution for $E^* \mid B^*$:

$$\mathbb{P}(E^* = e^* \mid E^* \geq e^*, B^*) = \begin{cases} \eta_{B^*,1} & \text{if } e^* < L_{\text{chunyun}}, \\ \eta_{B^*,2} & \text{if } e^* \geq L_{\text{chunyun}}, \end{cases}$$

where $L_{\text{chunyun}} = 41$ corresponds to January 10, the start of the Chinese New Year travel season known as "chunyun". We assume $\eta_{0,i} = \eta_{W,i}$ and $\eta_{1,i} = \cdots = \eta_{L,i} = \eta_{V,i}$, for $i = 1, 2$.

Under these different modeling assumptions, likelihood functions for the parameters can be computed in the same way as in Section 3.2, with integrals replaced by finite sums. We omit the details here.

5.3 Prior distributions and details of the implementation

To simplify the computation, we assume the incubation period of COVID-19 is less than 30 days. It is common to use a unimodal distribution with a smooth density function to model the incubation period. We use the following prior distribution on $h^*(\cdot)$ to encourage smoothness and log-concavity:

$$\begin{aligned} \pi\big(h^*(0), \ldots, h^*(29)\big) \propto & \Big(\prod_{x^*=0}^{29} h^*(x^*)^{\mu \cdot h_0(x^*)-1} \Big) \\ & \times \exp\Big\{ \sum_{x^*=1}^{28} \big(2 \log h^*(x^*) - \log h^*(x^*-1) - \log h^*(x^*+1)\big)_- \Big\}. \end{aligned} \tag{25}$$

where $(\cdot)_-$ is the negative part function. The first part of the right hand side of (25) is proportional to the density of a Dirichlet distribution with concentration parameters $\{\mu \cdot h_0(0), \ldots, \mu \cdot h_0(29)\}$. We choose $h_0(\cdot)$ to be a discretization of Gamma$(9, 1.5)$, whose tail probability of ≥ 14 days is less than 0.01. The second part of the right hand side of (25) is an exponential tilt which penalizes lack of log-concavity.

We put uninformative priors on other parameters in the model:

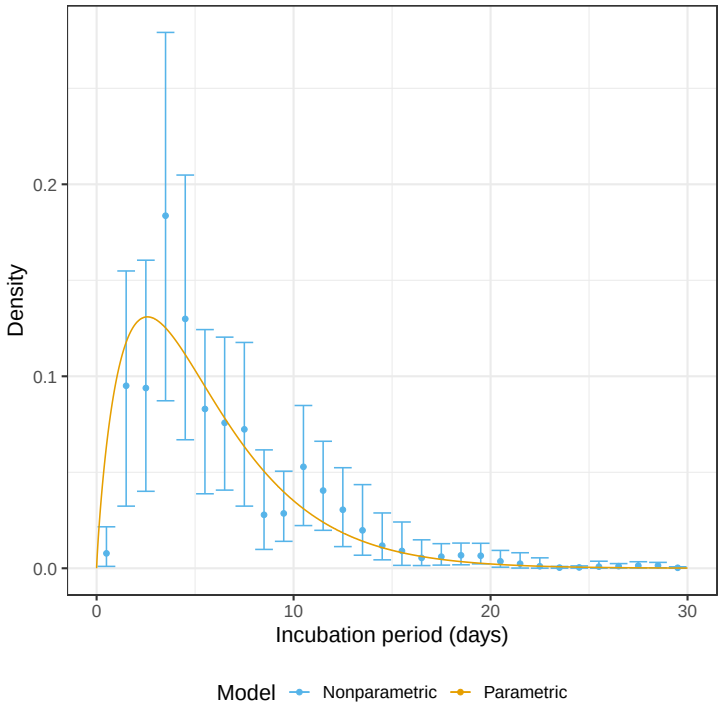
$$r_1 \sim \text{Exp}(1),\ r_2 \sim \text{N}(0, 4), \kappa \sim \text{Unif}(0, 1),\ \lambda_W, \lambda_V \sim \text{Unif}(0, 1/L).$$

Note that r_2 is allowed to be negative (exponential decrease after January 20). For the model with a geometric distribution for $E^* \mid B^*$, we put Unif$(0, 1)$ priors on $\eta_{W,1}, \eta_{W,2}, \eta_{V,1}, \eta_{V,2}$.

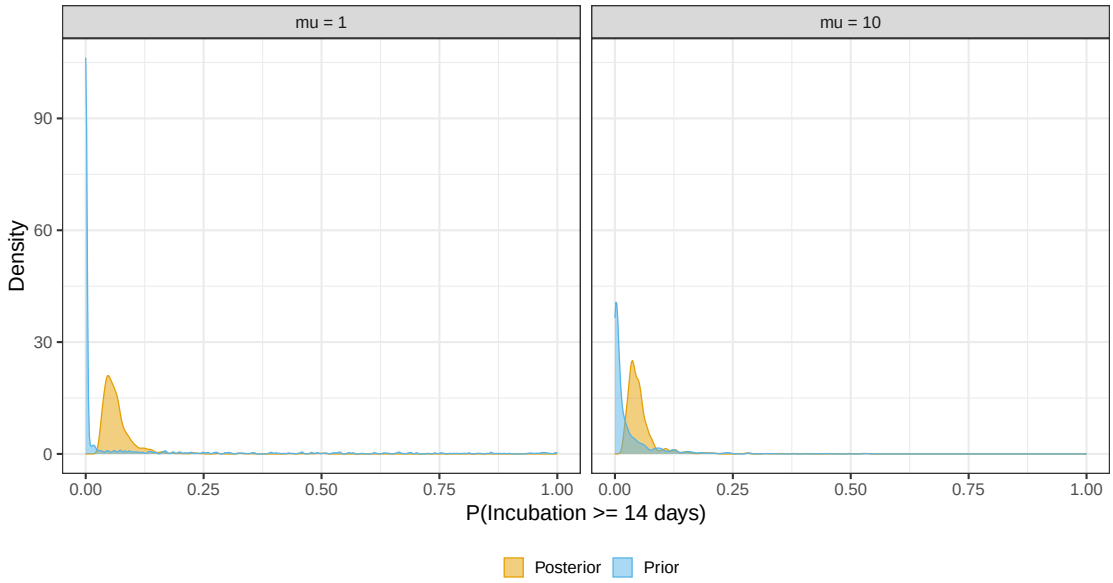
A random walk Metropolis–Hastings algorithm targeting the posterior distribution of $h^*(\cdot)$ and r_1, r_2 was implemented using the TensorFlow Probability library in Python [6]. We simulated chains of $10,000$ steps, discarding a burn-in period of 30%. The convergence of the sampler was assessed by simulating 8 parallel Markov chains with initial values overdispersed with respect to the target distribution, and computing the potential of scale reduction factor [10]. In every case, the statistic was confidently below 1.1.

5.4 Results of Bayesian nonparametric inference

Table 3 reports the results of the Bayesian nonparametric inference in 7 different scenarios. Overall, they are not too dissimilar to the results of the parametric model in Table 2. In particular, the bulk of nonparametrically estimated incubation period is quite similar to the Gamma distribution fitted in Section 3 (Figure 7a).



(a) A comparison of the estimated parametric and nonparametric incubation period distributions. Orange curve: Gamma distribution with median = 4.5 and 95% quantile = 13.4. Blue error bars: Posterior mean and 95% credible interval, computed in the model with a two-stage exponential growth, and geometric $E^* \mid B^*$, with $\mu = 10$.



(b) A comparison of the prior and posterior $\mathbb{P}(\text{Incubation} \geq 14 \text{ days})$ in the first two simulation scenarios of Table 3. Two panels corresponds to two choices for the prior distribution (25) of $h^*(\cdot)$.

Figure 7: An illustration of the nonparametric Bayesian fit to the incubation period distribution.

Sample		All	All	Shenzhen	Wuhan residents	All except Xinyang	All	All
Growth		r_1	r_1	r_1	r_1	r_1	r_1, r_2	r_1, r_2
$E^* \mid B^*$		Uniform	Uniform	Uniform	Uniform	Uniform	Uniform	Geometric
μ		$\mu = 1$	$\mu = 10$	$\mu = 1$	$\mu = 1$	$\mu = 10$	$\mu = 1$	$\mu = 1$
Doubling days for r_1		2.4 (2.2–2.6)	2.4 (2.2–2.6)	2.5 (2.2–2.9)	2.3 (2.1–2.5)	2.4 (2.2–2.6)	2.2 (2.0–2.4)	2.2 (2.0–2.5)
r_2 (Growth in Jan. 21–23)		–	–	–	–	–	.03 (-.20–.22)	-.11 (-.38–.12)
Incubation period	Mean	5.6 (5.1–6.2)	5.5 (5.1–5.9)	4.7 (3.8–6.2)	5.6 (5.1–6.2)	5.2 (4.7–5.6)	5.7 (5.1–6.3)	5.8 (5.3–6.5)
	$\mathbb{P}(\geq 7)$	.32 (.25–.39)	.31 (.25–.38)	.23 (.13–.35)	.30 (.24–.37)	.26 (.22–.31)	.06 (.03–.08)	.06 (.03–.09)
	$\mathbb{P}(\geq 10)$	.19 (.14–.24)	.18 (.14–.23)	.11 (.06–.19)	.20 (.13–.26)	.15 (.12–.21)	.19 (.15–.25)	.20 (.15–.27)
	$\mathbb{P}(\geq 14)$	.05 (.03–.08)	.04 (.02–.06)	.05 (.01–.12)	.05 (.03–.08)	.05 (.03–.07)	.06 (.03–.08)	.06 (.03–.09)
	$\mathbb{P}(\geq 21)$	.01 (.00–.01)	.00 (.00–.00)	.02 (.00–.08)	.01 (.00–.03)	.01 (.00–.03)	.01 (.01–.02)	.01 (.00–.03)

Table 3: Results of the nonparametric Bayesian inference where we do not impose a parametric form for the distribution of the incubation period. As sensitivity analyses, we also vary the study sample, model for the epidemic growth, distribution of E^* given B^*, and the hyperprior parameter μ. Numbers reported in the table are posterior means and 95% credible intervals (in brackets).

However, there is still a noticeable difference between the parametric and nonparametric fits. Without restricting the tail to follow that of a Gamma distribution, the estimated tail probabilities are higher than in Table 2. The posterior mean for $\mathbb{P}(S^* - T^* \geq 14 \text{ days})$ exceeds 0.04 in all scenarios, even when we exclude the cases confirmed in Xinyang who seemed to have longer incubation periods in Table 2. Moreover, prior and posterior distributions of $\mathbb{P}(S^* - T^* \geq 14 \text{ days})$ show a large discrepancy (Figure 7b), indicating that the posterior estimates of the tail probabilities are driven by the data instead of the prior. Taken together, this shows that the the restriction to the family of Gamma distributions (Assumption 3) very likely lead to under-estimation of the tail incubation period, and the probability of an incubation period of at least 14 days may be as large as 5%.

6 Discussion

In this article, we have proposed the generative BETS model for four key epidemiological events: beginning of exposure, end of exposure, time of transmission, and time of symptom onset. Under parametric models, we have derived the sample inclusion probability for exported cases and used it to correct for selection bias in the likelihood functions. Across different subsamples and modeling assumptions, the initial epidemic doubling time for COVID-19 in Wuhan was consistently estimated to be between 2 to 2.5 days. Our nonparametric Bayesian analysis suggests that the parametric fit likely under-estimated the tail of the incubation period, and among all the COVID-19 patients who develop symptoms, 5% of them could develop the symptoms at least 14 days after contracting the pathogenic virus. These estimates are pertinent to public health policies for the COVID-19 pandemic.

In addition to constructing a generative model and deriving the likelihood functions from first principles, we have also exposed the dangers of selection bias in some early epidemiological analyses of COVID-19. We find that the severity of selection bias was startling. This highlights the lesson that, when evaluating epidemiological and other real data studies, data quality and methodical considerations of selection bias are often much more important than data quantity and specific parametric models. This is especially important in high-stake decisions like the ones for the ongoing pandemic.

A key epidemiological parameter we decided not to study in this article is the basic reproduction number, commonly denoted by R_0. Intuitively, R_0 is the expected number of secondary infections produced by a typical case in a population where everyone is susceptible. In early outbreak analysis, R_0 can be estimated from the epidemic growth exponent r by $R_0 = 1/M(-r)$ [19], where $M(\cdot)$ is the moment generating function for the distribution of the serial interval (time between successive cases in a chain of transmission). Several studies have attempted to estimate the serial interval of COVID-19 in Wuhan by using observed pairs of infector-infectees [13, 15, 7]. The reported point estimate of the mean serial interval ranging from 4.0 [7] to 7.5 days [13]. However, for most COVID-19 cases it is impossible to ascertain the infector, so these early estimates of the serial interval could be severely biased by sample selection just like the early estimates of epidemic growth and incubation period as seen in Section 4.

Our findings in this article should be viewed together with limitations of our methodology. Although the contact tracing for travelers from Wuhan was intensive in the locations included in our dataset, some degree of under-ascertainment of Wuhan-exported cases is perhaps inevitable. There is also ambiguity about where some COVID-19 cases were infected if they both had stayed in Wuhan and were exposed to other confirmed cases after their stay. At the core of the BETS model, it is assumed that the disease transmission and progression are independent of traveling. This assumption is necessary to extend the conclusions from a "shadow" of the epidemic (Wuhan-exported cases)

to the center of the outbreak, but it can be violated if, for example, some people canceled travel plans due to feeling sick. It is also possible that the population of travelers is not representative of the general population in a meaningful way. Nevertheless, we believe these limitations are minor compared to ignoring the selection bias. We hope the generality of our model also makes it extensible in further studies of the current pandemic and other outbreaks in the future.

Acknowledgement

We thank Rajen Shah, Yachong Yang, Cindy Chen, Yang Chen, Dylan Small, Michael Levy, Hera He, Zilu Zhou, Yunjin Choi, James Robins, Marc Lipsitch, Andrew Rosenfeld for their helpful suggestions.

References

[1] Michael Gove: rate of coronavirus infection in UK doubling every three to four days – video. https://www.theguardian.com/world/video/2020/mar/27/michael-gove-rate-of-coronavirus-infection-in-uk-doubling-every-three-to-four-days-video. Retrieved: April 15, 2020.

[2] COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). https://coronavirus.jhu.edu/map.html. Retrieved: April 15, 2020.

[3] Boris Johnson coronavirus speech transcript: UK PM tells UK to avoid non-essential travel & contact. https://www.rev.com/blog/transcripts/boris-johnson-coronavirus-speech-transcript-uk-pm-tells-uk-to-avoid-non-essential-travel-contact. Retrieved: April 15, 2020.

[4] WHO statement regarding cluster of pneumonia cases in Wuhan, China. https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china, 2020. Retrieved: April 15, 2020.

[5] Jantien A Backer, Don Klinkenberg, and Jacco Wallinga. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance*, 25(5):2000062, 2020. doi: 10.2807/1560-7917.ES.2020.25.5.2000062.

[6] Joshua V Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A Saurous. Tensorflow distributions. *arXiv preprint arXiv:1711.10604*, 2017.

[7] Zhanwei Du, Xiaoke Xu, Ye Wu, Lin Wang, Benjamin J Cowling, and Lauren Ancel Meyers. The serial interval of covid-19 from publicly reported confirmed cases. *Emerging Infectious Diseases*, 26(6), 2020.

[8] Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994.

[9] Centers for Disease Control and Prevention. Interim clinical guidance for management of patients with confirmed coronavirus disease (covid-19). https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html. Retrieved: April 15, 2020.

[10] Andrew Gelman, Donald B Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

[11] Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of Internal Medicine*, 2020.

[12] Steffen L Lauritzen. *Graphical Models*, volume 17. Clarendon Press, 1996.

[13] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 2020.

[14] Natalie M Linton, Tetsuro Kobayashi, Yichi Yang, Katsuma Hayashi, Andrei R Akhmetzhanov, Sung-mok Jung, Baoyin Yuan, Ryo Kinoshita, and Hiroshi Nishiura. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *Journal of Clinical Medicine*, 9(2):E538, 2020.

[15] Hiroshi Nishiura, Natalie M Linton, and Andrei R Akhmetzhanov. Serial interval of novel coronavirus (covid-19) infections. *International Journal of Infectious Diseases*, 2020.

[16] Jonathan M Read, Jessica RE Bridgen, Derek AT Cummings, Antonia Ho, and Chris 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`.

[17] Nicholas G. Reich, Justin Lessler, Derek A. T. Cummings, and Ron Brookmeyer. Estimating incubation period distributions with coarse data. *Statistics in Medicine*, 28(22):2769–2784, 2009. doi: 10.1002/sim.3659.

[18] S Sanche, YT Lin, C Xu, E Romero-Severson, N Hengartner, and R Ke. High contagiousness and rapid spread of Severe Acute Respiratory Syndrome Coronavirus 2. *Emerging Infectious Diseases*, 26(7), 2020.

[19] Jacco Wallinga and Marc 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, 2007.

[20] Joseph T Wu, Kathy Leung, and Gabriel M Leung. Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in Wuhan, China: a modelling study. *Lancet*, 395(10225):689–697, 2020.

[21] Qingyuan Zhao, Yang Chen, and Dylan 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`.

A Technical proofs

A.1 Derivation of Lemma 1

Using eqs. (3) and (4), it is straightforward to show that

$$\begin{aligned}
&\mathbb{P}((B,E,T,S)\in\mathcal{D}\mid B=b,E=e)\\
=&\mathbb{P}(b\le T\le e,T\le S<\infty\mid B=b,E=e)\\
=&\int_{t\in(b,e)} f_T(t\mid b,e)\int_{s\in(t,\infty)} f_S(s\mid b,e,t)\,ds\,dt\\
=&\int_{t\in(b,e)} f_T(t\mid b,e)\Big\{\int_{s\in(t,\infty)}\nu\cdot h(s-t)\,ds\Big\}\,dt\\
=&\int_{t\in(b,e)} f_T(t\mid b,e)\cdot\nu\,dt\\
=&\nu[G(e)-G(b)].
\end{aligned}$$

A.2 Derivation of Lemma 2

By Assumption 3, $G_{\kappa,r}(t)=\int_{-\infty}^{t} g_{\kappa,r}(s)\,ds=(\kappa/r)\exp(rt)$. Thus for $b>0$, we have

$$\begin{aligned}
&\mathbb{P}((b,E,T,S)\in\mathcal{D}\mid B=b)\\
=&\nu\int_{e\in(b,L)} f_E(e\mid b)\left[G_{\kappa,r}(e)-G_{\kappa,r}(b)\right]de\\
=&\nu\int_b^L \lambda_V\,(\kappa/r)\left\{\exp(re)-\exp(rb)\right\}de\\
=&\frac{\lambda_V\kappa\nu}{r}\Big[\frac{1}{r}\big(\exp(rL)-\exp(rb)\big)-(L-b)\exp(rb)\Big]\\
=&\frac{\lambda_V\kappa\nu}{r^2}\exp(rL)-\frac{\lambda_V\kappa}{r}(r^{-1}+L-b)\exp(rb)\\
=&\frac{\lambda_V\kappa\nu}{r^2}\exp(rL)\Big[1-(1+r(L-b))\exp(-r(L-b))\Big].
\end{aligned}\tag{26}$$

For $b=0$, we can replace λ_V in the above equation by λ_W.

The idea is that, if rL is much larger than 1 (in our preliminary analysis $rL\approx 0.25\times 54=13.5$), then

$$\text{Right hand side of (26)}\approx\frac{\nu\lambda_W\kappa}{r^2}\exp(rL)\text{ when }b=0.$$

Using this approximation, we obtain

$$
\begin{aligned}
&\mathbb{P}((B,E,T,S)\in\mathcal{D})\\
=&\int_{0\le b<L}\mathbb{P}((b,E,T,S)\in\mathcal{D}\mid B=b)\,f_B(b)\,db\\
=&\,\mathbb{P}(B=0)\cdot\mathbb{P}((b,E,T,S)\in\mathcal{D}\mid B=0)+\int_{0<b<L}\mathbb{P}((b,E,T,S)\in\mathcal{D}\mid B=b)\,f_B(b)\,db\\
\approx&\frac{(1-\pi)\lambda_W\kappa\nu}{r^2}\exp(rL)+\int_{0<b<L}\mathbb{P}((b,E,T,S)\in\mathcal{D}\mid B=b)\,f_B(b)\,db\\
=&\frac{(1-\pi)\lambda_W\kappa\nu}{r^2}\exp(rL)+\pi\int_0^L\frac{1}{L}\frac{\lambda_V\kappa\nu}{r^2}\exp(rL)\Big[1-(1+r(L-b))\exp(-r(L-b))\Big]\,db\\
=&\frac{(1-\pi)\lambda_W\kappa\nu}{r^2}\exp(rL)+\frac{\pi\lambda_V\kappa\nu}{r^2}\exp(rL)-\frac{\pi}{L}\frac{\lambda_V\kappa\nu}{r^2}\exp(rL)\underbrace{\int_0^L\Big[(1+r(L-b))\exp(-r(L-b))\Big]}_{A_1}\,db\\
\approx&\frac{\kappa\exp(rL)\nu}{r^2}\Big[(1-\pi)\lambda_W+\pi\lambda_V(1-2/(rL))\Big].
\end{aligned}
$$

In the last step we used the approximation $e^{rL}\gg 1+rL$:

$$A_1=\int_0^L(1+rx)\exp(-rx)\,dx=-\frac{\exp(-rx)(rx+2)}{r}\Big|_{x=0}^{x=L}=\frac{2}{r}-\frac{\exp(-rL)(rL+2)}{L}\approx\frac{2}{r}.$$

Therefore, the conditional density is given by

$$
\begin{aligned}
f(b,e,t,s\mid\mathcal{D})\approx&\frac{[(1-\pi)\lambda_W 1_{\{b=0\}}+(\pi/L)\lambda_V 1_{\{b>0\}}]\cdot\kappa\exp(rt)\cdot\nu h(s-t)}{r^{-2}\kappa\exp(rL)\nu\Big[(1-\pi)\lambda_W+\pi\lambda_V(1-2/(rL))\Big]}\\
=&r^2\cdot\frac{[(1-\pi)\lambda_W 1_{\{b=0\}}+(\pi/L)\lambda_V 1_{\{b>0\}}]\cdot\exp(rt)}{\Big[(1-\pi)\lambda_W+\pi\lambda_V(1-2/(rL))\Big]\cdot\exp(rL)}\cdot h(s-t)\\
=&r^2\cdot\frac{[1_{\{b=0\}}+(\rho/L)1_{\{b>0\}}]\cdot\exp(rt)}{\Big[1+\rho(1-2/(rL))\Big]\cdot\exp(rL)}\cdot h(s-t),
\end{aligned}
$$

where $\rho=(\lambda_V/\lambda_W)\pi/(1-\pi)$.

A.3 Derivation of Proposition 1

The following Lemma is useful to marginalize over T when the incubation period follows a Gamma(α,β) distribution:

Lemma 3. *For any $r>0$ and $b\le e\le s$,*

$$\int_b^{\min(s,e)}\exp(rt)\,h_{\alpha,\beta}(s-t)\,dt=\Big(\frac{\beta}{\beta+r}\Big)^{\alpha}\exp(rs)\left[H_{\alpha,\beta+r}(s-b)-H_{\alpha,\beta+r}((s-e)_+)\right].$$

Proof. By a change of variables,

$$\begin{aligned}
&\int_b^{\min(s,e)} \exp(rt)\, h_{\alpha,\beta}(s-t)\, dt \\
=&\int_b^{\min(s,e)} \exp(rt)\, \frac{\beta^\alpha}{\Gamma(\alpha)}(s-t)^{\alpha-1}\exp\{-\beta(s-t)\}\, dt \\
=&\Big(\frac{\beta}{\beta+r}\Big)^\alpha \exp(rs)\int_b^{\min(s,e)} \frac{(\beta+r)^\alpha}{\Gamma(\alpha)}(s-t)^{\alpha-1}\exp\{-(\beta+r)(s-t)\}\, dt \\
=&\Big(\frac{\beta}{\beta+r}\Big)^\alpha \exp(rs)\left[H_{\alpha,\beta+r}(s-b) - H_{\alpha,\beta+r}((s-e)_+)\right].
\end{aligned}$$

□

The time of contraction T is not observed. Should it be observed, the full data unconditional likelihood is given by

$$\begin{aligned}
&L_{\text{uncond}}\big(\rho, r, h(\cdot); \boldsymbol{T}\big) \\
=&\prod_{i=1}^n f(B_i, E_i, T_i, S_i \mid (B_i, E_i, T_i, S_i) \in \mathcal{D}) \\
\approx& r^{2n}\cdot \prod_{i=1}^n \frac{1_{\{B_i=0\}} + (\rho/L)1_{\{B_i>0\}}}{1+\rho(1-2/(rL))} \cdot \prod_{i=1}^n \underbrace{1_{\{B_i\le T_i\le \min(E_i,S_i)\}}\cdot \exp(r(T_i-L))\cdot h(S_i-T_i)}_{A_{2,i}}.
\end{aligned}$$

If we assume $h(\cdot)$ is the density of a Gamma distribution:

$$h(x) = h_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x) \quad (x>0),$$

then we can marginalize over T_i using Lemma 3:

$$\int A_{2,i}\, dT_i = \exp\big\{r(S_i-L)\big\}\Big(\frac{\beta}{\beta+r}\Big)^\alpha \cdot \big[H_{\alpha,\beta+r}(S_i-B_i) - H_{\alpha,\beta+r}((S_i-E_i)_+)\big],$$

In conclusion, the unconditional observed data likelihood is given by

$$\begin{aligned}
L_{\text{uncond}}(\rho, r, \alpha, \beta) \approx r^{2n}\Big(\frac{\beta}{\beta+r}\Big)^{n\alpha}\cdot \prod_{i=1}^n \Bigg\{ &\frac{1_{\{B_i=0\}} + (\rho/L)1_{\{B_i>0\}}}{1+\rho(1-2/(rL))} \\
&\times \exp\big\{r(S_i-L)\big\}\big[H_{\alpha,\beta+r}(S_i-B_i) - H_{\alpha,\beta+r}((S_i-E_i)_+)\big]\Bigg\}.
\end{aligned}$$

The conditional observed data likelihood can be derived in the same way. Details are omitted.

A.4 Derivation of Proposition 2

By integrating the conditional density (17) over (b, e, s), the marginal distribution of T conditional on $(B, E, T, S) \in \mathcal{D}$ is given by

$$\begin{aligned}
f_T(t \mid \mathcal{D}) &= \int_{(b,e,t,s)\in\mathcal{D}} f(b, e, t, s \mid \mathcal{D})\, db\, de\, ds \\
&\approx \int_0^t \int_t^L \int_t^\infty r^2 \cdot \frac{[1_{\{b=0\}} + (\rho/L)1_{\{b>0\}}] \cdot \exp(rt)}{\Big[1 + \rho(1 - 2/(rL))\Big] \cdot \exp(rL)} \cdot h(s-t)\, ds\, de\, db \\
&= \int_0^t \int_t^L r^2 \cdot \frac{[1_{\{b=0\}} + (\rho/L)1_{\{b>0\}}] \cdot \exp(rt)}{\Big[1 + \rho(1 - 2/(rL))\Big] \cdot \exp(rL)}\, de\, db \\
&= \int_0^t (L-t) r^2 \cdot \frac{[1_{\{b=0\}} + (\rho/L)1_{\{b>0\}}] \cdot \exp(rt)}{\Big[1 + \rho(1 - 2/(rL))\Big] \cdot \exp(rL)}\, db \\
&= r^2(L-t) \cdot \frac{(1 + (\rho t/L)) \cdot \exp(rt)}{\Big[1 + \rho(1 - 2/(rL))\Big] \cdot \exp(rL)} \\
&\propto (L-t)(1 + (\rho t/L)) \exp(rt) \\
&\propto\!\!\propto (L-t) \exp(rt).
\end{aligned}$$

Assumption 2 says that the distribution of the symptom onset S only depends the time of transmission T, that is $S \perp\!\!\!\perp B, S \mid T$. Therefore the marginal distribution of S in Wuhan-exported cases is given by convolving the distribution of T with the distribution of the incubation period $S - T$:

$$f_S(s \mid \mathcal{D}) = \int_0^{\min(L,s)} f_T(t \mid \mathcal{D}) h(t-s)\, dt$$

Under the parametric assumption that $S - T$ follows a Gamma distribution (Assumption 4), we have (let $x = s - t$)

$$\begin{aligned}
f_S(s \mid \mathcal{D}) &\propto\!\!\propto \int_0^{\min(L,s)} (L-t) \exp(rt) \cdot (s-t)^{\alpha-1} \exp\{-\beta(s-t)\}\, dt \\
&= \exp(rs) \cdot \int_0^{\min(L,s)} [(L-s) + (s-t)](s-t)^{\alpha-1} \exp\{-(\beta+r)(s-t)\}\, dt \\
&= \exp(rs) \cdot \int_{(s-L)_+}^{s} [(L-s)x^{\alpha-1} + x^\alpha] \exp\{-(\beta+r)x\}\, dx \\
&= \exp(rs) \cdot \Big\{ (L-s) \frac{\Gamma(\alpha)}{(\beta+r)^\alpha} \big[H_{\alpha,\beta+r}(s) - H_{\alpha,\beta+r}((s-L)_+)\big] \\
&\qquad + \frac{\Gamma(\alpha+1)}{(\beta+r)^{\alpha+1}} \big[H_{\alpha+1,\beta+r}(s) - H_{\alpha+1,\beta+r}((s-L)_+)\big] \Big\} \\
&= \exp(rs) \cdot \Big\{ (L-s) \big[H_{\alpha,\beta+r}(s) - H_{\alpha,\beta+r}((s-L)_+)\big] \\
&\qquad + \frac{\alpha}{\beta+r} \big[H_{\alpha+1,\beta+r}(s) - H_{\alpha+1,\beta+r}((s-L)_+)\big] \Big\} \\
&\approx \exp(rs) \cdot \Big\{ (L-s)[1 - H_{\alpha,\beta+r}((s-L)_+)] + \frac{\alpha}{\beta+r}[1 - H_{\alpha+1,\beta+r}((s-L)_+)] \Big\}
\end{aligned}$$

A.5 Derivation of Proposition 3

Let $e_- = \min(e, M)$. Then under Assumptions 1 and 2, for $(b, e, t, s) \in \mathcal{D}$ and $s \le M$,

$$\begin{aligned} f_{T,S}(t, s \mid b, e, \mathcal{D}, S \le M) &= \frac{f_{T,S}(t, s \mid b, e, \mathcal{D})}{\int\int f_{T,S}(t, s \mid b, e, \mathcal{D})\, ds\, dt} \\ &= \frac{g(t)h(s-t)}{\int_b^{e_-} g(t) \int_t^M h(s-t)\, ds\, dt} \\ &= \frac{g(t)h(s-t)}{\int_b^{e_-} g(t)H(M-t)\, dt}, \end{aligned} \tag{27}$$

where $H(s) = \int_0^s h(x)\, dx$ is the distribution function of the incubation period. Assuming $g(t) = \kappa \exp(rt)$ and using integration by parts, for $r \neq 0$,

$$\begin{aligned} \int_b^{e_-} g(t)H(M-t)\, dt =& \kappa \int_b^{e_-} \exp(rt)H(M-t)\, dt \\ =& \frac{\kappa}{r} \int_b^{e_-} H(M-t)\, d\exp(rt) \\ =& \frac{\kappa}{r}\Big[\exp(rt)H(M-t)\Big|_{t=b}^{t=e_-} + \int_b^{e_-} \exp(rt)h(M-t)\, dt\Big]. \end{aligned}$$

By using $h(\cdot) = h_{\alpha,\beta}(\cdot)$ and using Lemma 3, we have

$$\int_b^{e_-} g(t)H_{\alpha,\beta}(M-t)\, dt = \frac{\kappa}{r}\Big[\exp(rt)H_{\alpha,\beta}(M-t) - \Big(\frac{\beta}{\beta+r}\Big)^\alpha \exp(rM)H_{\alpha,\beta+r}(M-t)\Big]\Big|_{t=b}^{t=e_-}.$$

Now we integrate t in (27) from b to e_- and get

$$\begin{aligned} & f_S(s \mid b, e, \mathcal{D}, S \le M) \\ =& \frac{r\Big(\frac{\beta}{\beta+r}\Big)^\alpha \exp(rs)\left[H_{\alpha,\beta+r}(s-b) - H_{\alpha,\beta+r}((s-e)_+)\right]}{\Big[\exp(rt)H_{\alpha,\beta}(M-t) - \Big(\frac{\beta}{\beta+r}\Big)^\alpha \exp(rM)H_{\alpha,\beta+r}(M-t)\Big]\Big|_{t=b}^{t=e_-}}. \end{aligned}$$

For $r = 0$, using integration by parts,

$$\begin{aligned} \int_b^{e_-} g(t)H_{\alpha,\beta}(M-t)\, dt =& \kappa \int_{(M-e)_+}^{M-b} H_{\alpha,\beta}(x)\, dx \\ =& \kappa\Big[xH_{\alpha,\beta}(x)\Big|_{x=(M-e)_+}^{x=M-b} - \int_{(M-e)_+}^{M-b} xh_{\alpha,\beta}(x)\, dx\Big] \\ =& \kappa\Big[xH_{\alpha,\beta}(x)\Big|_{x=(M-e)_+}^{x=M-b} - \int_{(M-e)_+}^{M-b} x \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)\, dx\Big] \\ =& \kappa\Big[xH_{\alpha,\beta}(x)\Big|_{x=(M-e)_+}^{x=M-b} - \frac{\alpha}{\beta} \int_{(M-e)_+}^{M-b} \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)} x^\alpha \exp(-\beta x)\, dx\Big] \\ =& \kappa\Big[xH_{\alpha,\beta}(x) - \frac{\alpha}{\beta} H_{\alpha+1,\beta}(x)\Big]\Big|_{x=(M-e)_+}^{x=M-b}. \end{aligned}$$

We can similarly integrate t out and obtain the full data likelihood. Details are ommited.