BETS: The dangers of selection bias in early analyses of the COVID-19 pandemic

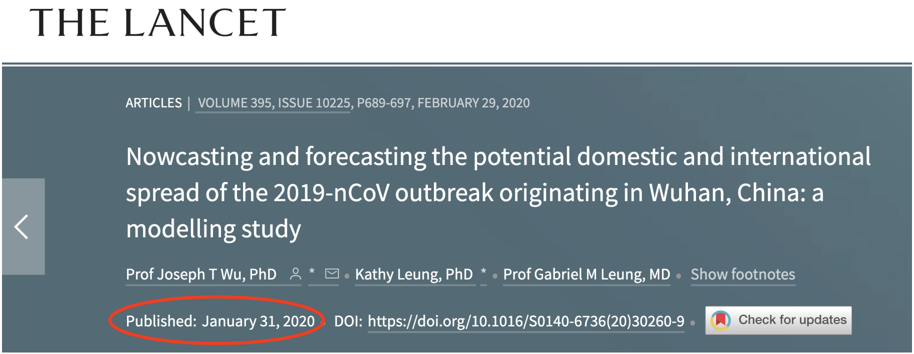
Nianqiao (Phyllis) Ju
5-th year Ph.D. student
Dept. of Statistics, Harvard University

joint work with Q. Zhao, S. Bacallado & R. Shah
at Statistical Laboratory, University of Cambridge

December 21st, 2020 @ CMStatistics 2020

Manuscript: arXiv:2004.07743
(forthcoming in *The Annals of Applied Statistics*)
Slides: https://nianqiaoju.github.io

A puzzling comparison



THE LANCET

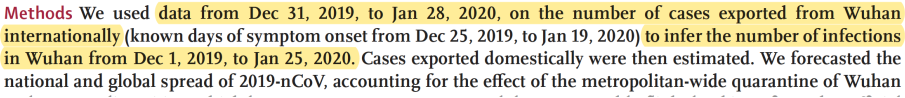
ARTICLES | VOLUME 395, ISSUE 10225, P689-697, FEBRUARY 29, 2020

Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study

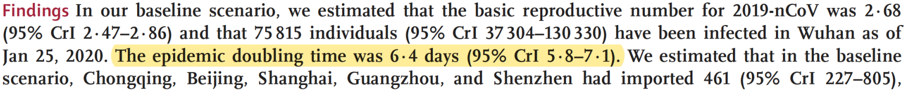
Prof Joseph T Wu, PhD * • Kathy Leung, PhD * • Prof Gabriel M Leung, MD • Show footnotes

Published: January 31, 2020 • DOI: https://doi.org/10.1016/S0140-6736(20)30260-9 •

Check for updates

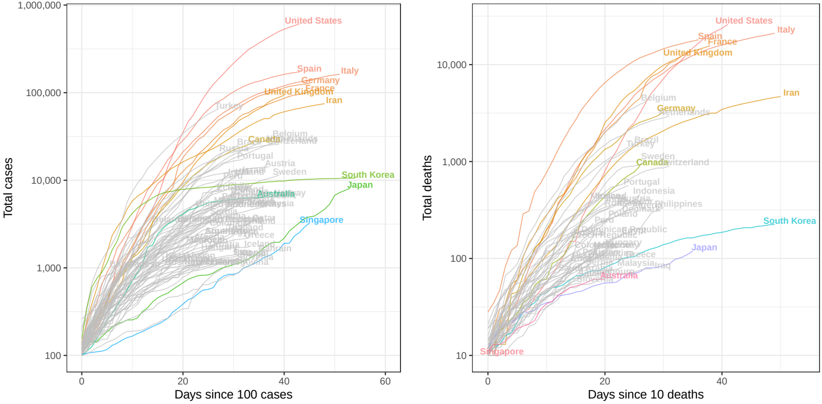


Methods We used data from Dec 31, 2019, to Jan 28, 2020, on the number of cases exported from Wuhan internationally (known days of symptom onset from Dec 25, 2019, to Jan 19, 2020) to infer the number of infections in Wuhan from Dec 1, 2019, to Jan 25, 2020. Cases exported domestically were then estimated. We forecasted the national and global spread of 2019-nCoV, accounting for the effect of the metropolitan-wide quarantine of Wuhan



Findings In our baseline scenario, we estimated that the basic reproductive number for 2019-nCoV was 2·68 (95% CrI 2·47–2·86) and that 75 815 individuals (95% CrI 37 304–130 330) have been infected in Wuhan as of Jan 25, 2020. The epidemic doubling time was 6·4 days (95% CrI 5·8–7·1). We estimated that in the baseline scenario, Chongqing, Beijing, Shanghai, Guangzhou, and Shenzhen had imported 461 (95% CrI 227–805),

Which one is correct?



In countries most hard hit by COVID-19, the total cases and deaths grew about 100 times in the first 20 days (doubling time: $20/\log_2(100) = 3.01$ days).

Similar data, different inference?

The Lancet study ignored the lockdown of Wuhan on Jan. 23rd.

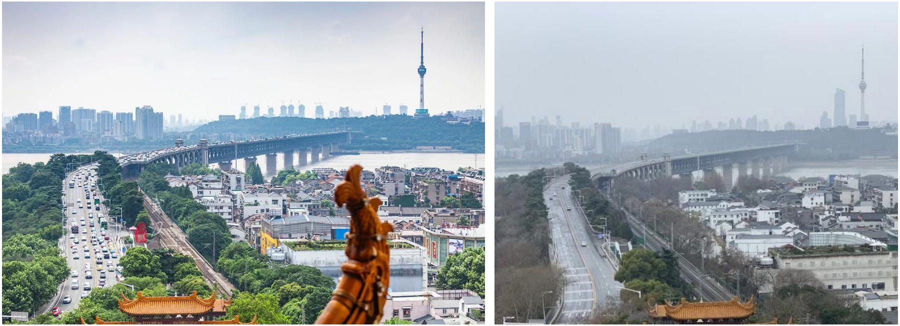


Figure: (left) before the lockdown (picture taken on Sept. 7, 2019) and (right) after the lockdown.

Selection bias

1. under-ascertainment bias: symptomatic patients did not seek healthcare or could not be diagnosed.
2. non-random selection bias: cases included in the study are not representative of the population.
3. travel ban: ignoring the travel ban leads to under-estimation of epidemic growth.
4. epidemic growth: patients were more likely to be infected towards the end of their exposure period.
5. right-truncation: cases confirmed after a certain time are excluded from the dataset.

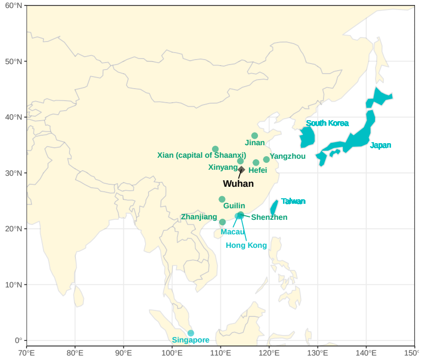
Selection bias recap

selection bias	epidemic growth	incubation period
under-ascertainment bias		
non-random selection bias		
travel ban	under-estimation	
epidemic growth		over-estimation
right-truncation		under-estimation

Keys to avoid the selection bias:

1. Carefully design the study and adhere to the sample inclusion criterion.
2. Start from a generative model and derive liklihood functions that adjust for sample selection.

Data collection



- 14 locations where the local health agencies published full case reports.
- 1,460 COVID-19 cases that were confirmed by February 29 for locations in mainland China (February 15 for international locations).

Overview of the dataset

available at https://github.com/qingyuanzhao/bets.covid19

Column name	Description	Example	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% NA)
Age	Age	63	Mean=45.6, IQR=[34, 57]
Known Contact	Known epidemiological contact?	[✓ Yes]/No	84.7%/15.3%
Cluster	Relationship with other cases	Husband of HongKong-04	32.1%known
Outside	**Transmitted outside Wuhan?**	Yes/[✓ Likely]/No	58.5%/7.7%/33.8%
Begin Wuhan	**Begin of stay in Wuhan (B)**	30-Nov	
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
Symptom	**Date of symptom onset (S)**	23-Jan	9.0% NA
Initial	Date of first medical visit	23-Jan	6.5% NA
Confirmed	Date confirmed	24-Jan	

Discerning Wuhan-exported cases

We obtained 378 cases exported from Wuhan that satisfy the following criteria:

- The case had stayed in Wuhan before January 23.
- The case had no recorded contact with other confirmed cases, or had the earliest symptom onset in their (family) cluster, or showed symptoms before they left Wuhan.
- The case did not have missing symptom onset.
- The case arrived at the location where they were diagnosed before January 24.

The principle is to only include cases as Wuhan-exported that pass a **“beyond a reasonable doubt”** test.

A generative model

Four crucial epidemiological events

- B: Beginning of stay in Wuhan;
- E: End of stay in Wuhan;
- T: Time of transmission (unobserved);
- S: Time of symptom onset.

Below we will:

- Define the support $\mathcal{P}$ of (B, E, T, S) for the **Wuhan-exposed** population;
- Construct a generative model for (B, E, T, S);
- Define the sample selection set $\mathcal{D}$ corresponds to **Wuhan-exported** cases;
- Derive likelihood functions to adjust for the sample selection.

Wuhan-exposed population $\mathcal{P}$

Intuitively, $\mathcal{P} =$ All people who stayed in Wuhan between December 1, 2019 (time 0) and January 24, 2020 (time L, the lockdown).

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

Under the following conventions.

- $B = 0$: Started their stay in Wuhan before time 0.
- $E = \infty$: Did not arrive in the 14 locations we are considering before time L. (We do not differentiate between people who stayed in Wuhan or went to a different location).
- $T = \infty$: Were not infected during their stay in Wuhan. (We do not differentiate between infection outside Wuhan and never infected.)
- $S = \infty$: Did not show symptoms of COVID-19 (never infected or asymptomatic).

Wuhan-exported cases

The event of observing **Wuhan-exported cases** can be written as

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

Recall that **Wuhan-exposed population** is

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

The Wuhan-exported cases have

1. $B \leq T \leq E$,
2. $E \leq L$,
3. $S < \infty$.

A generative BETS model

$$f(b,e,t,s) = \underbrace{f_B(b) \cdot f_E(e \mid b)}_{\text{travel}} \cdot \underbrace{f_T(t \mid b,e)}_{\text{disease transmission}} \cdot \underbrace{f_S(s \mid b,e,t)}_{\text{disease progression}} .$$

The BETS model makes two basic assumptions:

Assumption 1: Disease transmission independent of travel

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

Here $g(\cdot)$ models the **epidemic growth** in Wuhan before the lockdown.

A generative BETS model

$$f(b,e,t,s) = \underbrace{f_B(b)\cdot f_E(e \mid b)}_{\text{travel}} \cdot \underbrace{f_T(t \mid b,e)}_{\text{disease transmission}} \cdot \underbrace{f_S(s \mid b,e,t)}_{\text{disease progression}} .$$

Assumption 2: Disease progression independent of travel

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

Here $h(\cdot)$ is the density of the **incubation period** $S - T$ (for symptomatic cases).

Parametric assumptions

To ease the interpretation and simply the likelihood functions, we assume:

Assumption 3: Exponential growth

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

Assumption 4: Gamma-distributed incubation period

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

Which likelihood function?

For a moment, let's pretend the time of transmission T is observed.

✗ Sample from $\mathcal{P}$

$$\prod_{i=1}^{n} f(B_i, E_i, T_i, S_i)$$

✓ Sample from $\mathcal{D}$ (Unconditional likelihood)

$$\prod_{i=1}^{n} f(B_i, E_i, T_i, S_i \mid \mathcal{D}), \text{ where } f(b, e, t, s \mid \mathcal{D}) \triangleq \frac{f(b, e, t, s) \cdot 1_{\{(b,e,t,s) \in \mathcal{D}\}}}{\mathbb{P}((B, E, T, S) \in \mathcal{D})}.$$

✓ Sample from $\mathcal{D}$ (Conditional likelihood)

$$\prod_{i=1}^{n} f(T_i, S_i \mid B_i, E_i, \mathcal{D}), \text{ where } f(t, s \mid b, e, \mathcal{D}) \triangleq \frac{f(t, s \mid B = b, E = e) \cdot 1_{\{(b,e,t,s) \in \mathcal{D}\}}}{\mathbb{P}((B, E, T, S) \in \mathcal{D} \mid B = b, E = e)}.$$

Unobserved T

In reality, the time of transmission T is unobserved. We can use the marginal likelihood:

Unconditional likelihood

$$L_{\text{uncond}}(\theta) = \prod_{i=1}^{n} \int f\left(B_i, E_i, t, S_i \mid \mathcal{D}\right) dt,$$

where $\theta = (f_B(\cdot), f_E(\cdot \mid \cdot), g(\cdot), h(\cdot))$.

Conditional likelihood

$$L_{\text{cond}}(\theta) = \prod_{i=1}^{n} \int f\left(t, S_i \mid B_i, E_i, \mathcal{D}\right) dt,$$

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

Results

Location	Sample size	Doubling time (in days)	Incubation period Median	Incubation period 95% quantile
		Conditional likelihood		
China - Hefei	34	2.1 (1.2–3.7)	4.3 (2.9–6.0)	12.0 (9.1–17.3)
China - Shaanxi	53	1.7 (1.0–2.8)	4.5 (3.1–6.2)	14.6 (11.5–19.8)
China - Shenzhen	129	2.2 (1.7–3.0)	3.5 (2.8–4.3)	11.2 (9.5–13.6)
China - Xinyang	74	2.3 (1.5–3.5)	6.8 (5.4–8.2)	16.4 (13.8–20.1)
China - Other	42	2.0 (1.1–3.4)	5.1 (3.6–6.7)	12.3 (9.8–16.4)
International	46	2.1 (1.4–3.4)	3.8 (2.5–5.3)	10.9 (8.4–15.1)
All locations	378	2.1 (1.8–2.5)	4.5 (4.0–5.0)	13.4 (12.2–14.8)
		Unconditional likelihood		
China - Hefei	34	1.8 (1.4–2.4)	4.1 (2.8–5.5)	11.9 (9.0–17.2)
China - Shaanxi	53	2.5 (2.0–3.1)	5.3 (3.9–6.8)	15.0 (12.0–20.0)
China - Shenzhen	129	2.4 (2.1–2.8)	3.6 (2.9–4.3)	11.3 (9.6–13.7)
China - Xinyang	74	2.4 (2.0–2.9)	6.8 (5.6–8.1)	16.4 (13.9–20.2)
China - Other	42	2.1 (1.7–2.8)	5.3 (4.0–6.6)	12.4 (10.0–16.4)
International	46	2.0 (1.6–2.6)	3.7 (2.5–5.0)	10.8 (8.4–15.1)
All locations	378	2.3 (2.1–2.5)	4.6 (4.1–5.1)	13.5 (12.3–14.9)

(Point estimates obtained by MLE. Confidence intervals obtained by inverting LRT.)

What happened in the Lancet study?

Wu et al. uses SEIR model and assumes density of S in $\mathcal{P}$ is

$$f(s \mid \mathcal{P}) \underset{\sim}{\propto} \exp(rs), \text{ for } s \leq L.$$

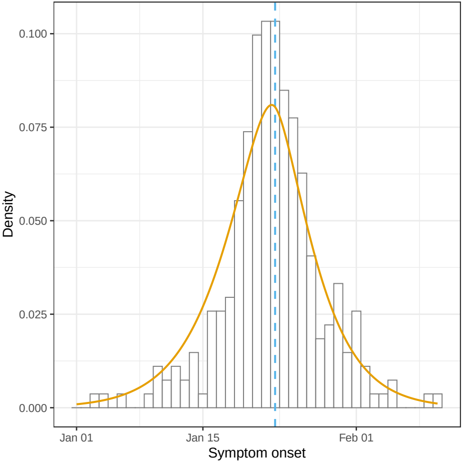
Under our model and some reasonable assumptions, density of S in $\mathcal{D}$ is

$$f(t \mid \mathcal{D}, B = 0) \underset{\sim}{\propto} \exp(rt)\,(L - t)\,1_{\{t \leq L\}},$$

and

$$f_S(s \mid \mathcal{D}, B = 0) \underset{\sim}{\propto} \exp(rs)\left(L + \frac{\alpha}{\beta + r} - s\right), \text{ for } s \leq L.$$

Effect of traval ban



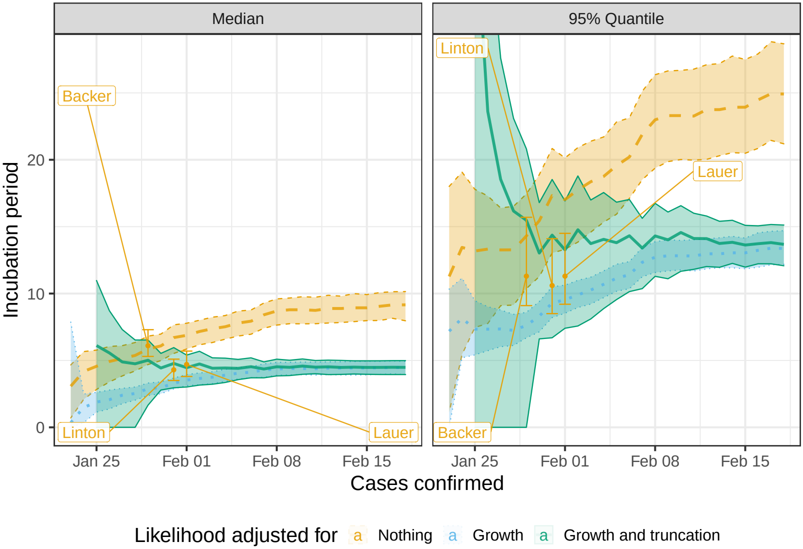
- Histogram: Density of the symptom onset of the Wuhan-resident cases;
- Orange curve: Theoretical fit $f_S(s \mid \mathcal{D}, B = 0)$ using MLE of (r, α, β).
- Blue dashed line: January 23, 2020 (time L).

Selection bias recap

bias	remedy
under-ascertainment bias	
non-random selection bias	$\mathcal{D}$ v.s. $\mathcal{P}$
travel ban	$f(\cdot \mid \cdot, \mathcal{D})$
epidemic growth	$L_{cond}(r, \alpha, \beta)$
right-truncation	$L_{cond}(r, \alpha, \beta; M)$

Keys to avoid the selection bias:

1. Carefully design the study and adhere to the sample inclusion criterion.
2. Start from a generative model and derive liklihood functions that adjust for sample selection.



Ignore epidemic growth $\Longrightarrow$ Overestimate incubation period.

Ignore right-truncation $\Longrightarrow$ Underestimate incubation period.

Conclusions

Conclusions about COVID-19

- Initial doubling time in Wuhan: 2–2.5 days.
- Median incubation period: about 4 days.
- Proportion of incubation period at least 14 days: about 5%.

Our study has many limitations:

- Reported symptom onset could be inaccurate.
- Some degree of under-ascertainment is perhaps inevitable.
- Discerning Wuhan-exported cases is not black-and-white.
- Assumptions 1 & 2 (independence of travel and disease) could be violated.

Conclusions

Compelling evidence for selection bias in early studies

(i) Under-ascertainment.

(ii) Non-random sample selection.

(iii) Travel ban.

(iv) Epidemic growth.

(v) Right-truncation.

Don't make uncalculated BETS

1. Carefully design the study and adhere to the sample inclusion criterion.
2. Base statistical inference on first principles.

Final lesson

As statisticians and data practitioners, we are blessed by the wealth of data in this digital age. But let's not forget an important lesson:

$$\textbf{Data Quality + Better Design} \gg \textbf{Data Quantity + Better Model}$$

Manuscript: arXiv:2004.07743
(forthcoming in *The Annals of Applied Statistics*)
Slides: `https://nianqiaoju.github.io`