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

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

May 5, 2020 @ YSPH Biostatistics Seminar

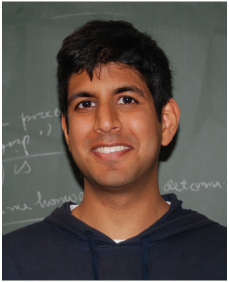
Manuscript: arXiv:2004.07743

Slides: `http://www.statslab.cam.ac.uk/~qz280/`.

# Collaborators

Nianqiao (Phyllis) Ju
PhD student at Harvard

Sergio Bacallado
Stats Lab, Cambridge



Rajen Shah
Stats Lab, Cambridge

## And many thanks to...

Cindy Chen, Yang Chen, Yunjin Choi, Hera He, Michael Levy, Marc Lipsitch, James Robins, Andrew Rosenfeld, Dylan Small, Yachong Yang, Zilu Zhou, and many other who have provided helpful suggestions.

# COVID-19 is personal for everyone


Me and my parents, all grew up in in Wuhan, China.
(September 7, 2019)

# Wuhan Lockdown (January 23, 2020)

Before the lockdown

After the lockdown

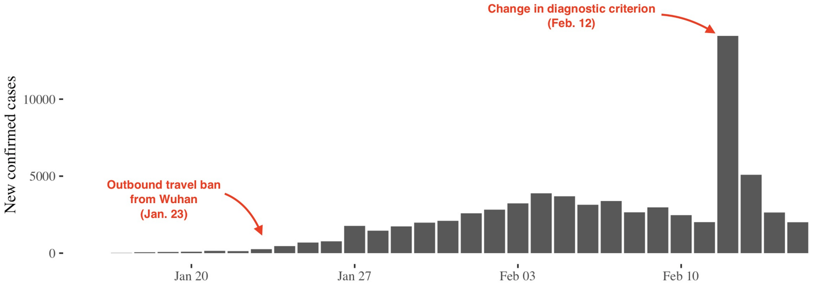
# The beginning of this project

- On January 29, I heard from my parents that a close relative was just diagnosed with "viral pneumonia". This prompted me to start looking into the data available at the time.
- **However, epidemiological data from Wuhan are very unreliable!**

## Some anecdotal evidence

- **Inadequate testing:** The relative of mine could not get a RT-PCR test till mid-February, when she was already recovering.
- **False negative test:** Her first test was negative. A few days later she was tested again and the result came back positive.
- **Insufficient contact tracing:** Her husband who also showed COVID symptoms quickly recovered and was never tested.

# Insufficient testing in Wuhan




A change of diagnostic criterion on February 12 led to a huge spike of cases.

## Solution: Using cases "exported" from Wuhan

This has two benefits:

1. Testing and contact tracing were intensive in other locations.
2. Detailed case reports (instead of mere case counts) are often available.

This design was first used by Neil Ferguson's team in Imperial College, who estimated on January 17 that there might be already over 1,700 cases in Wuhan.

# Our first analysis



Previous

**Analysis of the epidemic growth of the early 2019-nCoV outbreak using internationally confirmed cases**

Posted February 09, 2020.

Qingyuan Zhao, Yang Chen, Dylan S Small

**doi:** https://doi.org/10.1101/2020.02.06.20020941

**Download PDF**

Data/Code

**Methods:** We obtained information on the 46 coronavirus cases who traveled from Wuhan before January 23 and have been subsequently confirmed in Hong Kong, Japan, Korea, Macau, Singapore, and Taiwan as of February 5, 2020. Most cases have detailed travel history and disease progress. Compared to previous analyses, an important distinction is that we used this

**Results:** We found that our model provides good fit to the distribution of the infection time. Assuming the travel rate to the selected countries and regions is constant over the study period, we found that the epidemic was doubling in size every 2.9 days (95% credible interval [CrI], 2 days—4.1 days). Using previously reported serial interval for 2019-nCoV, the estimated basic

# 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

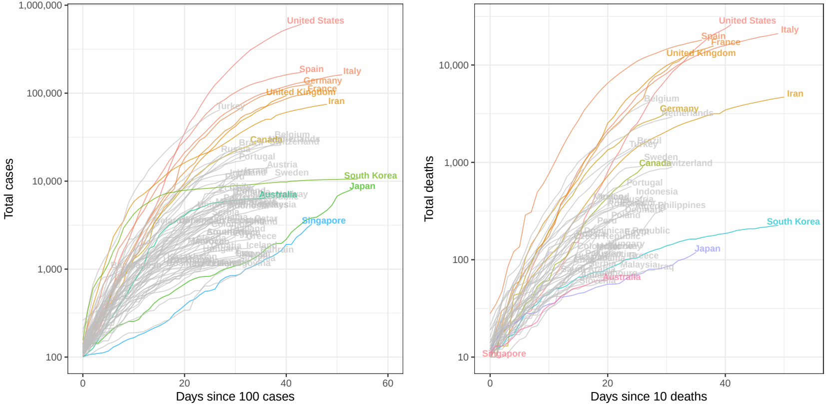
Prof Joseph T Wu, PhD * • Kathy Leung, PhD * • Prof Gabriel M Leung, MD • Show footnotes

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).

# How can the results be so different?

## Spoilers...

Similar data and model were used in these two studies, with one crucial difference:

**The Lancet study did not take into account the travel ban.**

REUTERS

Business Markets World Politics TV More

WORLD NEWS JANUARY 23, 2020 / 0:59 AM / 3 MONTHS AGO

**Wuhan lockdown 'unprecedented', shows commitment to contain virus: WHO representative in China**

# Rest of the talk

# Bias (i): Under-ascertainment

- This may occur if symptomatic patients did not seek healthcare or could not be diagnosed.
- **Susceptible studies:** All studies using cases confirmed when testing is insufficient.
- **Direction of bias:** Varied, depending on the pattern of under-ascertainment and parameter of interest.
- **Solution:** Use carefully considered and planned study designs.

# Bias (ii): Non-random sample selection

- Cases included in the study are not representative of the population.
- **Susceptible studies:** All studies, as detailed information of COVID-19 cases is sparse, but especially those without clear inclusion criteria.
- **Direction of bias:** Varied.
- **Solution:** Follow a protocol for data collection and exclude data that do not meet the sample inclusion criterion.

# Bias (iii): Travel ban

- Outbound travel from Wuhan was banned from January 23, 2020 to April 8, 2020.
- **Susceptible studies:** Studies that analyze cases exported from Wuhan.
- **Direction of bias:** Under-estimation of epidemic growth and infection-to-recovery time.
- **Solution:** Derive tailored likelihood functions to account for travel restrictions.

# Bias (iv): Epidemic growth

- Patients were more likely to be infected towards the end of their exposure period.
- **Susceptible studies:** Studies that treat infections as uniformly distributed over the exposure period.
- **Direction of bias:** Over-estimation of the incubation period.
- **Solution:** Derive tailored likelihood functions to account for epidemic growth.

# Bias (v): Right-truncation

- Cases confirmed after a certain time are excluded from the dataset.
- **Susceptible studies:** Studies that only use cases detected early in an epidemic.
- **Direction of bias:** Under-estimation of the incubation period.
- **Solution:**
  1. Collect all cases that meet a selection criterion, do not end data collection prematurely;
  2. Derive tailored likelihood functions to correct for right-truncation.

# Recap

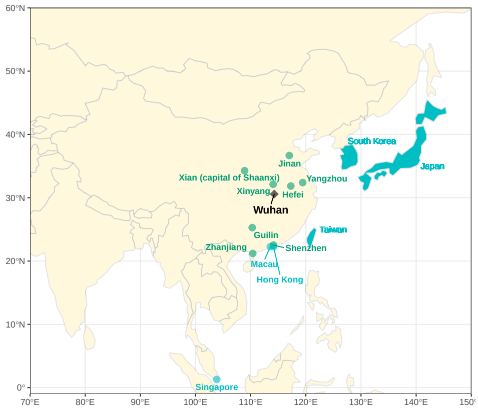
## Types of bias in COVID-19 analyses

(i) Under-ascertainment.
(ii) Non-random sample selection.
(iii) Travel ban.
(iv) Epidemic growth.
(v) Right-truncation.

## 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 likelihood 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

| 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[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<br>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 12am December 1, 2019 (time 0) and 12am January 24, 2020 (time $L$, the lockdown).

## 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).

Under these conventions.

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

# A generative BETS model

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

To allow extrapolation from Wuhan-exported sample to Wuhan-exposed population, 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.

## 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**

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

**Assumption 4: Gamma-distributed incubation period**

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

- The nuisance parameters $\nu$ (proportion of symptomatic cases) and $\kappa$ (baseline transmission) will be canceled in the likelihood function.
- Assumptions 3 & 4 will be relaxed in the Bayesian nonparametric analysis.

# Wuhan-exported cases

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

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

This makes three further restrictions on $\mathcal{P}$:

1. $B \leq T \leq E$, because we only use cases who contracted the virus during their stay in Wuhan;
2. $E \leq L$, because the case can only be observed if they left Wuhan before the travel ban;
3. $S < \infty$, because we only consider COVID-19 cases who showed symptoms.

# 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)

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

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

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

# Unobserved $T$

In reality, the time of transmission $T$ is unobserved. We can either treat $T$ as a latent variable and use e.g. an EM algorithm, or use the **integrated likelihood**:

## Unconditional likelihood

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

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

## Conditional likelihood

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

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

The conditional likelihood is less efficient because it does not use information in $f(b, e \mid \mathcal{D})$; but it is robust to misspecifying the travel models $f_B(\cdot), f_E(\cdot \mid \cdot)$.

# Conditional likelihood function

## Proposition

Under Assumptions 1–4,

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

where $H_{\alpha,\beta}(\cdot)$ is the CDF of Gamma$(\alpha,\beta)$ and $(\cdot)_+ = \max(\cdot,0)$ is the positive part function.

- Does not depend on $\nu$ (proportion of symptomatic cases) and $\kappa$ (baseline transmission).
- When $r=0$, reduces to the likelihood function in Reich et al. (2009) *Statistics in Medicine*, 28:2769–2784.

# Unconditional likelihood function

## Assumption 5: Stable travel

1. Beginning of stay $B$ follows a uniform distribution given $0 < B \leq L$.
2. End of stay $E$ follows a uniform distribution from $B$ to $L$ (with different rates for Wuhan residents and Wuhan visitors).

## Proposition

Under Assumptions 1–5 and suitable approximations,

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

where $\rho$ is a traveling parameter (capturing the different traveling patterns between Wuhan residents and visitors).

# Results

| Location | Sample size | $\rho$ | Doubling time (in days) | Incubation period Median | Incubation period 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) |

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

# Conclusions from the parametric model

- The initial doubling time in Wuhan is between 2 to 2.5 days.
- The median incubation period is around 4 days.
- The 95% quantile of the incubation period is between 11 to 15 days.

# 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),

# What happened?

Wu et al. used a modified SEIR (Susceptible-Exposed-Infectious-Recovered) model to account for traveling. But they **did not consider the travel ban**.

## ✗ Density of $S$ in $\mathcal{P}$

It is reasonable to assume incidence of symptom onset is growing exponentially in **Wuhan-exposed population** $\mathcal{P}$:

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

But we are sampling from the **Wuhan-exported cases** $\mathcal{D}$.

## ✓ Density of $S$ in $\mathcal{D}$

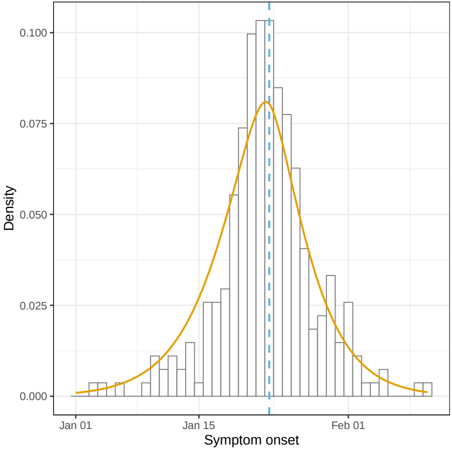
Under Assumptions 1–5 and reasonable approximations,

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

We can further derive the theoretical $f_S(s \mid \mathcal{D}, B = 0)$; in particular,

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

# Illustration of the selection bias (iii)




- 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 $(r, \alpha, \beta)$.
- Blue dashed line: January 23, 2020 (time $L$).

# Bias (iv): Epidemic growth

- Patients were more likely to be infected towards the end of their exposure period.
- **Susceptible studies:** Studies that treat infections as uniformly distributed over the exposure period.
- **Direction of bias:** Over-estimation of the incubation period.
- **Solution:** **Use the likelihood $L_{\textbf{cond}}(r, \alpha, \beta)$ instead of $L_{\textbf{cond}}(0, \alpha, \beta)$.**

# Bias (v): Right-truncation

- Cases confirmed after a certain time are excluded from the dataset.
- **Susceptible studies:** Studies that only use cases detected early in an epidemic.
- **Direction of bias:** Under-estimation of the incubation period.
- **Solution:** **Derive the likelihood with the additional conditioning event $S \leq M$.**

## Likelihood function adjusted for right-truncation

- Under Assumptions 1 & 2,

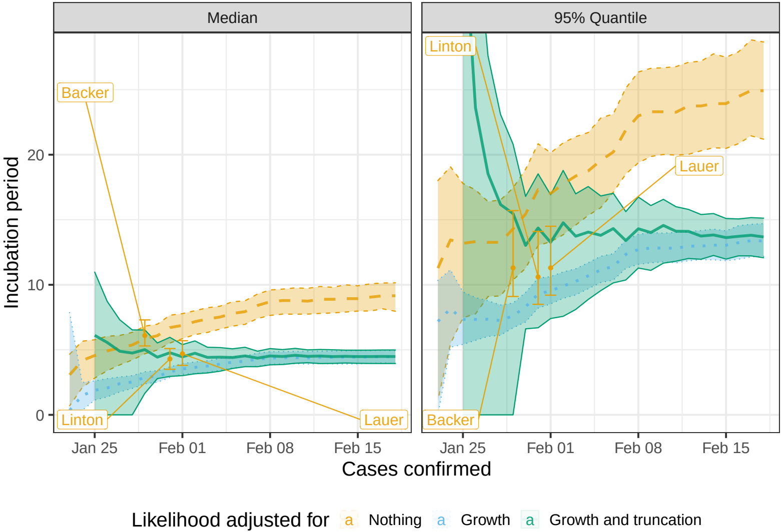
$$f_{T,S}(t, s \mid b, e, \mathcal{D}, S \leq M) = \frac{g(t)h(s-t)}{\int_b^{\max(e,s)} g(t)H(M-t)\,dt},$$

  where $H(\cdot)$ is the CDF of $h(\cdot)$.
- Closed-form expression for $L_{\text{cond,trunc}}(r, \alpha, \beta; M)$ can further be obtained under Assumptions 3 & 4 using integration by parts.

# Illustration of the selection bias (iv) and (v)

## An experiment

- For each day between January 23 and February 18, obtain the subset of cases confirmed by that day.
- Fit the parametric BETS model by using one of the following likelihoods:
  1. **Adjusted for nothing:** $L_{\text{cond}}(0, \alpha, \beta)$ (likelihood function in Reich et al. (2009) used in other studies).
  2. **Adjusted for growth:** $L_{\text{cond}}(r, \alpha, \beta)$.
  3. **Adjusted for growth and right-truncation:** $L_{\text{cond,trunc}}(r, \alpha, \beta; M)$.
- Obtain point estimates by MLE and CIs by nonparametric Bootstrap.
- Compare with previous studies:
  1. Backer, J. A. et al. *Eurosurveillance*, 25(5), 2020. PubMed: 32046819.
  2. Lauer, S. A. et al. *Annals of Internal Medicine*, 2020. PubMed: 32150748.
  3. Linton, N. M. et al. *Journal of Clinical Medicine*, 9(2), 2020. PubMed: 32079150.




Ignore epidemic growth $\Longrightarrow$ Overestimate incubation period.
Ignore right-truncation $\Longrightarrow$ Underestimate incubation period.

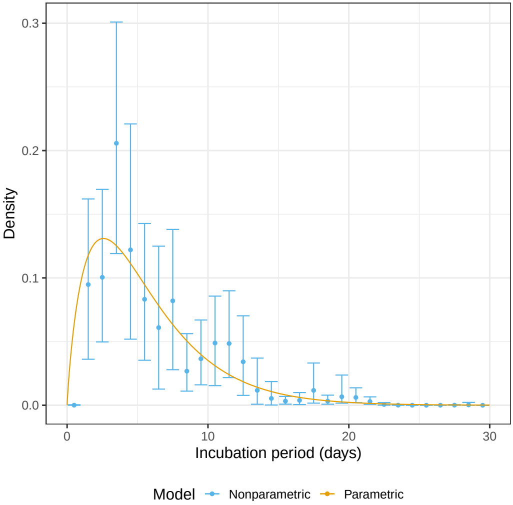
# Nonparametric model

- Assumption 4 (Gamma-distributed incubation period) may overly restrict the shape of the tails.
- We further consider Bayesian nonparametric inference for the incubation period.

## Implementation details

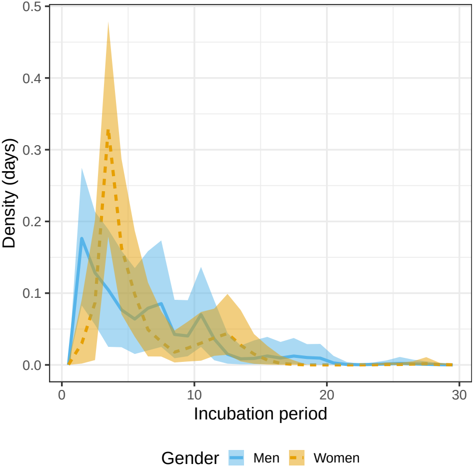
- First discretize the model: $B^* = \lceil B \rceil$, $E^* = \lceil E \rceil$, $T^* = \lceil T \rceil$, $S^* = \lceil S \rceil$.
- The extend the continuous-time model to the discrete time. Density $h(\cdot)$ of the incubation period becomes point masses: $h^*(0), h^*(1), \dots, h^*(29)$.
- A prior distribution is put on $h^*$ to encourage smoothness and log-concavity.
- Variations of Assumption 3 (exponential growth) and Assumption 5 (stable travel) are implemented to test sensitivity.

# Parametric vs. Nonparametric fit




Posterior estimate of $\mathbb{P}(S^* - T^* \geq 14)$ is about 5%, slightly higher than before.

# Gender-specific incubation period




- More men develop symptoms within 2 days of infection (physiology?).
- Men have heavier tail incubation period than women (behavior?).

# 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:**

**Data Quality + Better Design $\gg$ Data Quantity + Better Model**