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

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

Sep 8, 2021 @ RSS Discussion Meeting

Timeline of COVID-19

2019

Dec 1	Symptom onset of the index case.
Dec 31	First public message and alert.

2020

Jan 13	First confirmed case outside of China.
Feb 7	Deaths exceeds those of SARS.
Mar 11	Declared as a pandemic by the WHO.
Apr 1	1 million cases recorded.
Sep 10	1 million deaths recorded.
Dec 2	First approval of a COVID-19 vaccine.

2021

Apr 25	1 billion vaccine doses administered.
Jun 9	New naming system for variants of interest and concern.

Most pressing questions in Jan & Feb 2020

Question 1

Can COVID-19 be transmitted from human to human?

Question 2

How fast was COVID-19 growing?

In hindsight

- The answers are extremely obvious.
- Yet the correct conclusions weren't reached until they were extremely obvious.

Why?

Disclaimers

This presentation is retrospective

- Data are always limited at the beginning of a disease outbreak.

This presentation is biased by personal attachments

- Wuhan is my hometown.
- My aunt was a victim and had a long recovery from long COVID.
- I tried to warn about it (medRxiv:2020.02.06.20020941):

  Results: . . . The epidemic was doubling in size every 2.9 days (95% CrI 2 to 4.1) . . . The estimated basic reproduction number is 5.7 (95% CrI 3.4 to 9.2).

  Conclusions: . . . This indicates the 2019-nCoV could have been spreading faster than previous estimates.

Source of information

- I do not have more insider information than any other citizen of the Internet.
- Retrospect on Question 1 is largely based on reports by Chinese journalists.

Acronyms

COVID-19	COronaVIrus Disease 2019.
WHO	World Health Organization.
WMHC	Wuhan Municipal Health Commission.
CCDC	Chinese Center for Disease Control and Prevention.
SARS	Severe Acute Respiratory Syndrome.
PUE	Pneumonia of Unknown Etiology.
MERS	Middle East Respiratory Syndrome.

Question 1: Human-to-human transmissible?

Three expert groups were sent by CCDC to Wuhan on Dec 31, Jan 8, and Jan 8.

WMHC press releases/Expert conclusions

No clear evidence of human-to-human transmission. (Jan 5, 2020)

[The disease] can be prevented and contained. (Jan 9, 2020)

No clear evidence of human-to-human transmission . . . Although we cannot exclude the possibility of limited human-to-human transmission, the risk of sustained transmission is low. (Jan 14, 2020)

Novel coronavirus can certainly be transmitted from human to human. (Jan 20, 2020)

Table: Total number of confirmed cases in Wuhan by date of press release.

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

Evidence of high infectiousness

1. COVID-19's pathogen—SARS-CoV-2—is far from the first human coronavirus.
2. Genomic sequencing already identified a SARS-like coronavirus in late December, 2019.
3. Several hospitals in Wuhan were seeing rapidly increasing numbers of suspected cases in early and mid January, 2020.

In hindsight, there had been enough evidence by January 10, 2020 to conclude that the novel coronavirus was very likely to be highly infectious.

Why was the conclusion not reached sooner?

1. The authorities adopted a cautious and conservative approach towards infectious diseases.
   - For example, all unofficial laboratories were asked to destroy their samples and to not disclose their existing results on January 3, 2020.
2. Case definition was too strict.
   - Official case number once dropped from 59 (Jan 5) to 41 (Jan 11) and remained flat till Jan 17.
   - A controversial "white booklet" printed by WMHC made epidemiological exposure a necessary condition for PUE.
3. CCDC's surveillance system for PUE did not work as intended.
4. Lack of coordination and statistical expertise.
   - The first investigatory group only consisted of respiratory experts.
   - The second group included epidemiologists, but no reports on their efforts to gather additional evidence before concluding the outbreak "can be prevented and contained".
5. Lack of risk assessment.

Question 2: How fast was COVID-19 growing?

Initial quantitative studies focused on R_0 (basic reproductive number)

Basic reproductive number

- $R_0 \approx$ average number of infectees per infected person in the beginning of an outbreak.
- Estimated R_0 of SARS: 2 to 4; MERS: < 1; seasonal influenza: < 2.
- Important to determine the chance of a large outbreak and the "herd immunity" threshold.
- I realized much later that the precise definition of R_0 is model-dependent.

An arguably more useful metric for early outbreaks is the initial doubling time:

- Better captures the urgency;
- Is much easier to estimate.

Two initial studies

1. Q. Li, et al. Early transmission dynamics in Wuhan, China, of novel coronavirusinfected pneumonia. *New England Journal of Medicine*.
    - Published online: Jan 29, 2020
    - Estimated initial doubling time: **7.4 days (95% CI, 4.2 to 14)**;
    - Estimated R_0: **2.2 (95% CI, 1.4 to 3.9)**;
    - 13244 citations (Google Scholar, Aug 30, 2021).
2. J. T. Wu, et al. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *Lancet*.
    - Published online: Jan 31, 2020.
    - Estimated initial doubling time: **6.4 days (95% CI 5.8 to 7.1)**;
    - Estimated R_0: **2.68 (95% CI 2.47 to 2.86)**.
    - 3557 citations (Google Scholar, Aug 30, 20201).

Below: Some re-analyses

I failed to replicate their results using simple models.

First paper: Author description of statistical analysis

We estimated the epidemic growth rate by analyzing data on the cases with illness onset between December 10 and January 4 because we expected the proportion of infections identified would increase soon after the formal announcement of the outbreak in Wuhan on December 31.

We fitted a transmission model (formulated with the use of renewal equations) with zoonotic infections to onset dates that were not linked to the Huanan Seafood Wholesale Market, and we used this model to derive the epidemic growth rate, the epidemic doubling time, and the basic reproductive number (R_0).

- No particulars about the transmission model was mentioned besides the model was fitted using MATLAB.

First paper: Re-analysis

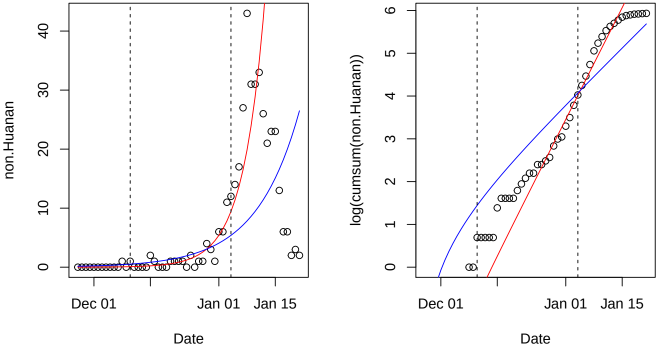


Figure: Initial epidemic curve in Wuhan and the fitted log-linear models (using the incidences between December 10 and January 4, dashed lines). The red curves correspond to an unrestricted fit; the blue curves correspond to the best fit assuming that the growth exponent correspond to a doubling time of 7.4 days.

- Initial doubling time by a Poisson log-linear model: **3.7 days (2.8 to 5.1)**.

Second paper: Author description of statistical analysis

78 exported cases from Wuhan to areas outside mainland China.

We used the following susceptible-exposed-infectious-recovered (SEIR) model to simulate the Wuhan epidemic

We assumed that ... international case exportation occurred according to a non-homogeneous process.

We estimated R_0 using Markov Chain Monte Carlo methods with Gibbs sampling and non-informative flat prior.

We estimated ... on the basis of ... confirmed cases ... whose symptom onset date had been reported to fall from Dec 25, 2019, to Jan 19, 2020. ... This end date was chosen to minimise the effect of lead time bias on case confirmation.

- Not mentioned: this criterion left them with **just 17 cases**.

Second paper: Re-analysis

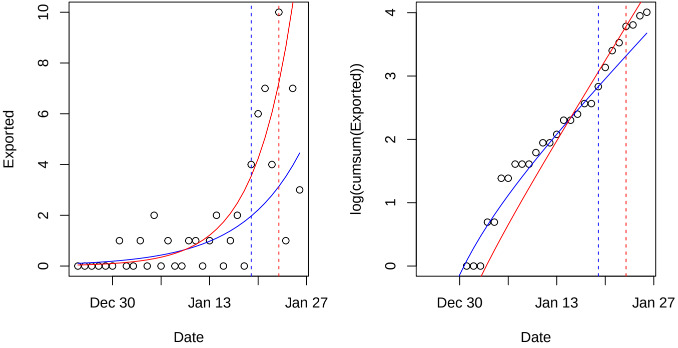


Figure: Initial epidemic curve for Wuhan-exported cases and the fitted log-linear models. The blue curves correspond to using the incidences between December 25 and January 19 (blue dashed lines). The red curves correspond to fitting the same model using data up to January 23 (red dashed lines).

- Initial doubling time by a Poisson log-linear model: **5.9 days (3.4 to 15.7)**.
- Confidence interval is much wider.
- Smaller point estimate (3.9 days) had the end date been extended.

Re-analyses: Conclusions

Compartmental (e.g. SEIR) models may be brittle for initial outbreaks analyses:

- Sensitive to choice of model parameters;
- Only uses incidence curves;
- Can easily overfit the data.

Beyond modelling epidemic curves

Incidence curve is only an insufficient summary of raw data

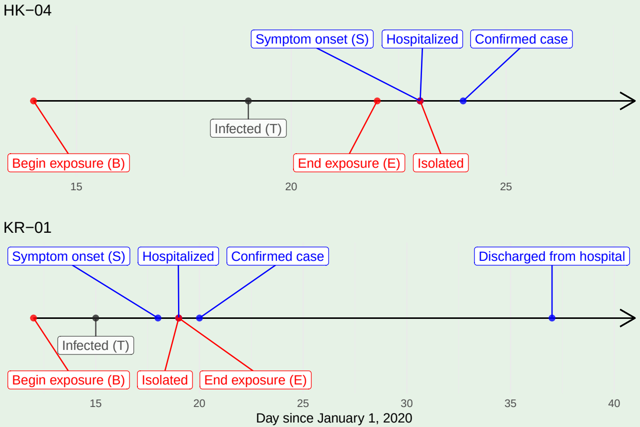


Figure: Timelines of two COVID-19 cases. The color indicates type of event.

A better statistical model

Key events in the trajectory

1. Begin of exposure, B;
2. End of exposure, E;
3. Onset of symptoms, S.

These events can help to infer the latent time of infection, T:

- Logical constraint: $B \leq T \leq E$;
- $S - T$ is the incubation period.

Sample selection: Travel quarantine of Wuhan on Jan 23

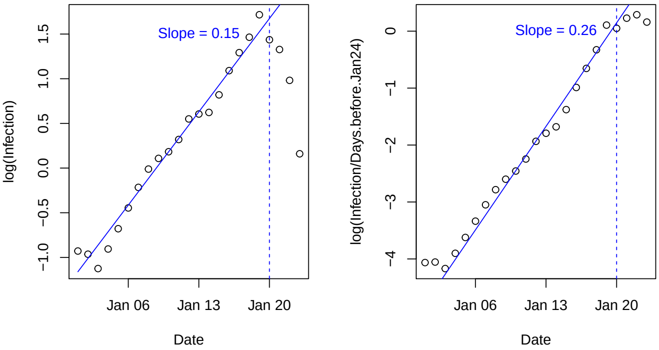
Instead of a simple exponential growth model for the density of T: $f_T(t) \propto e^{rt}$, a better model is

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

where L is the time of travel quarantine.

Re-analysis of exported cases

- Feb 2020 medRxiv preprint: Fit growth models to imputed infection times.



- This requires knowing the incubation period distribution.
- A better model can be found in Q Zhao, N Ju, S Bacallado, and R Shah. BETS: The dangers of selection bias in early analyses of the coronavirus disenase (COVID-19) pandemic. *Annals of Applied Statistics*.

Estimating R_0

Two ways to estimate R_0

1. R_0 as a derived parameter from compartmental models;
2. Through the formula $R = 1/M(-r)$:
   - $M(\cdot)$ is the moment generating function of the generation time;
   - r is the growth exponent.

- Both approaches often require estimating external parameters (e.g. distributions of the incubation period or generation time).
- Surprisingly (to statisticians), none of the early COVID-19 studies considered uncertainty in external parameters.

Large unaccounted uncertainty about R_0

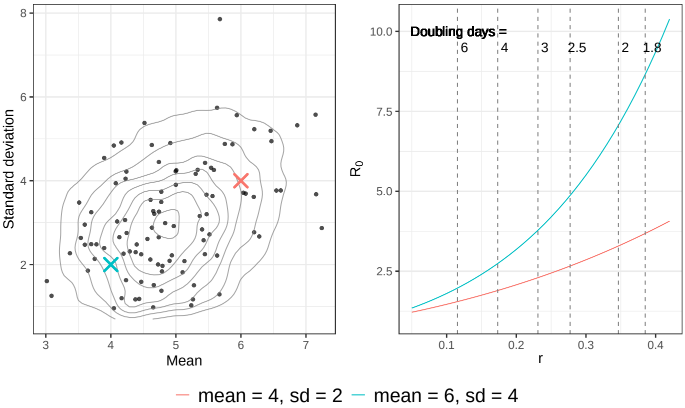


Figure: Estimated generation time and the implied basic reproductive number. The left panel shows 100 samples from the posterior distribution of the mean and standard deviation of the generation time in a replication analysis. The right panel shows the implied R_0 for different values of r and two generation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

Discussion

- Reliable data on COVID-19 were scarce, so the initial analyses are understandably imperfect.
- Nonetheless, had the investigators and decision makers been better prepared, some of the mistakes could have been easily avoided.

Lessons

1. Small data analysis is crucial but not easy;
2. Practitioners are still unfamiliar with basic statistical concepts;
3. Better research appraisal is in urgent need;
4. Right data are often more important than right analysis.