Estimation of incubation period distribution of COVID-19 using onset forward time: a novel cross-sectional and forward follow-up study

Jing Qin, Biostatistics research branch, NIAID.
Joint work with You, Lin, Hu, Yu and Xiao-Hua Zhou at Peking University.

Why incubation period estimation?

What is the incubation period? Based on Merriam Webster:
"The period between the infection of an individual by a pathogen and the manifestation of the illness or disease it causes."
In this definition, it has nothing to do with the PCR test and antibody test.
To prevent further spread of disease (for example COVID-19), a reliable estimation of incubation period helps to determine the idea length of quarantine.
We don't care a patient would develop symptoms on Day 3 or Day 10. The more important question is the tail estimation $P(X \geq 14)$? What is the proportion of patients develop disease after a 14-day quarantine?

Direct observations

Suppose we can observe incubation periods directly. Let $X_1, ...X_n$ be the observed incubation periods from a density $f(x)$.
Then mean and median can be estimated directly by sample mean and median. The 14 days tail probability can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} I(x_i \geq 14).$$

Nice and easy! No problem! No statistician is needed!

Problems

The most difficult problem in infectious disease study is to ascertain the infection onset time.

1. Recall bias.
2. Lack of ability to judge when was infected.
3. For those with long incubation periods, it is very difficult to tell when the disease was contracted. Some researchers may intentionally or unintentionally to truncate long incubation periods! However those patients with long incubation periods will have a contribution in the tail estimation.

Symptoms onset

The symptoms onset times are much easier to ascertain for those confirmed cases.
Current approach is to treat the exposure as an interval censoring problem.

Literature

Lauer, Grantz, Bi...., Justin Lessler (Johns Hopkins) in Annals of Internal Medicine reported that 97.5% COVID-19 patients would have symptoms onset within 11.5 days. It is almost absolutely impossible to have symptoms onset time after 14 days (probability less than 1%).
In a MedRxiv manuscript by Bi, ... Justin Lessler,..., they reported that 5% COVID-19 patients would develop symptoms after 14 days.
The results contradict each other even from the same research team!

Why?

1. Same problems as I mentioned in pervious slides. Direct observations on incubation period are difficult and not reliable.
2. Selection bias, or truncation problem!
3. Tail problem. It is very hard to estimate the tail based on small sample size.

Lauer et al. assembled $n = 181$ COVID-19 patients from China (73), Singapore (16), Japan (13)...
Bi et al. used $n = 183$ from a study in Shenzhen (near Hongkong).
To estimate the tail probability accurately, we need large sample size! $Y \sim Binorm(n, p)$, the 95% CI would be

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}.$$

The error margin is 0.042 if $n = 200, \hat{p} = 0.1$.

Implications

Lauer, Grantz, Bi...., Justin Lessler (Johns Hopkins) in Annals of Internal Medicine It is almost absolutely impossible to have symptoms onset time after 14 days (probability less than 1%. Bi, ... Justin Lessler,..., they reported (MedRxiv) that 5% COVID-19 patients would develop symptoms after 14 days.

	#Susceptible	Infection Rate	#Incidents$>$ 14 days
NYC	1000	10%	5
Utah	1000	1%	0

Take home message

The length of a quarantine period should be set carefully in regions with a severe epidemic.

Data

In the early stage of epidemic outbreak, Chinese government lockdown Wuhan in January 23, as well as almost the whole country.
A lot of asymptomatic people left Wuhan and then developed symptom somewhere else in the later follow up study.
12,953 confirmed cases were collected in Dr. Anderw Zhou's group (Peking University) based on daily reports from provincial and municipal health commissions in China.

Data

Confirmed cases	12,963
Dates of symptoms onset collected	6,345
Travel or residency history in Wuhan	3,168
Departure dates recorded	2,514
Departure dates and symptoms onset dates recorded	1,922
Departure between 19-23 Jan	1,211

Why only select cases who departed between 19-23 Jan? The lockdown policy was implemented strictly in China after January 23.

If someone left Wuhan too early, it is hard to know where and when this individual contracted the disease.

Trend

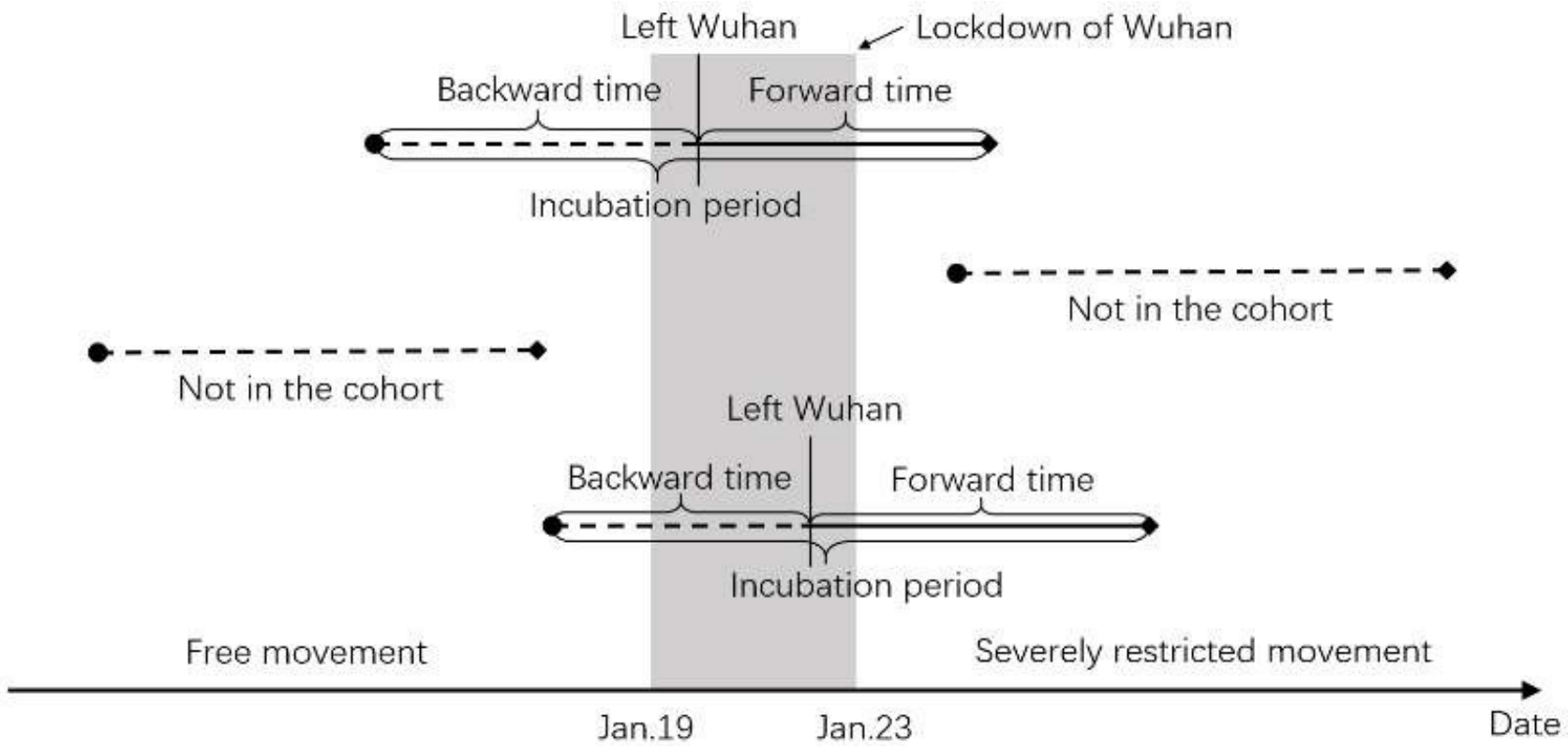

Figure: Illustration of our cross-sectional and forward follow-up study. Backward and incubation periods are not observed, while Wuhan departure and forward time are observed.

Conventional survival analysis does not work!

The incubation period X is calculated from contracting the disease in Wuhan to symptoms onset. What we observed V is asymptomatic on departure from Wuhan to symptoms onset, which can be treated as the censored version of X.

$$V = X - \textit{departure Wuhan time} > 0, \quad X > V.$$

If X has a density f and survival function $\bar{F}$, then the likelihood is

$$L = \prod_{i=1}^{n} \bar{F}(v_i)$$

In order to maximize L, $\bar{F}(v_i)$ must be 1 for all i!
Even a Weibull distribution is assumed,

$$L = \prod_{i=1}^{n} \exp\{-(\lambda v_i)^{\alpha}\}$$

Clearly $\lambda = 0$ is the MLE!

Renewal process

D. R. Cox, Renewal Theory (1964).
A renewal process is a sequence $\{X_i, i = 1, 2, ...\}$ of indep. and identically distributed random variables. $X_i \geq 0$.
$E(X) < \infty, V(X) < \infty$.

$$T_n = \sum_{i=1}^{n} X_i$$

is the n-th renewal takes place. Let

$$N(t) = \max\{n : T_n \leq t\}$$

be the number of renewals in $(0, t]$.

Forward time and backward time

The forward time is defined as

$$V(t) = T_{N(t)+1} - t,\ t > 0$$

the time between any given time t and the next epoch of the renewal process.
The backward time is defined as

$$A(t) = t - T_{N(t)}$$

the time between the last epoch of the renewal process to t.
The inter arrival interval is defined as

$$Y(t) = T_{N(t)+1} - T_{N(t)} = X_{N(t)+1} = A(t) + V(t)$$

Renewal process paradox

Suppose the incubation period X has a density $f(x)$ ($\bar{F}$ is survival). We may treat from contracting disease to symptoms onset as a renewal.

As $t \to \infty$, the observed $Y = A + V$ is a length biased version of X, where A is the time difference between sampling time (around January 23) and disease onset time, and V is calculated from sampling time to symptoms onset. A is called backward time and V is called forward time.

$$(A, V) \sim \frac{f(a+v)}{\mu},\ a, v > 0,\ \mu = \int_0^\infty tf(t)$$

$$Y \sim \frac{yf(y)}{\int yf(y)dy}$$

$$A \sim V \sim \frac{\bar{F}(v)}{\mu}$$

The key: For each individual, it is not necessarily $A = V$, however, A and V have the same distribution.

Maximum likelihood estimation

We have assumed the incubation period to be a Weibull distribution. The Weibull density is

$$f(x) = \alpha\lambda(\lambda x)^{\alpha-1}\exp\{-(x\lambda)^{\alpha}\}$$

The forward time

$$V \sim g(v) = \frac{\bar{F}(v)}{\mu} = \alpha\lambda\exp\{-(v\lambda)^{\alpha}\}/\Gamma(1/\alpha)$$

The likelihood is

$$\prod_{i=1}^{n} g(v_i, \alpha, \lambda)$$

where $n = 1211$,

Trend

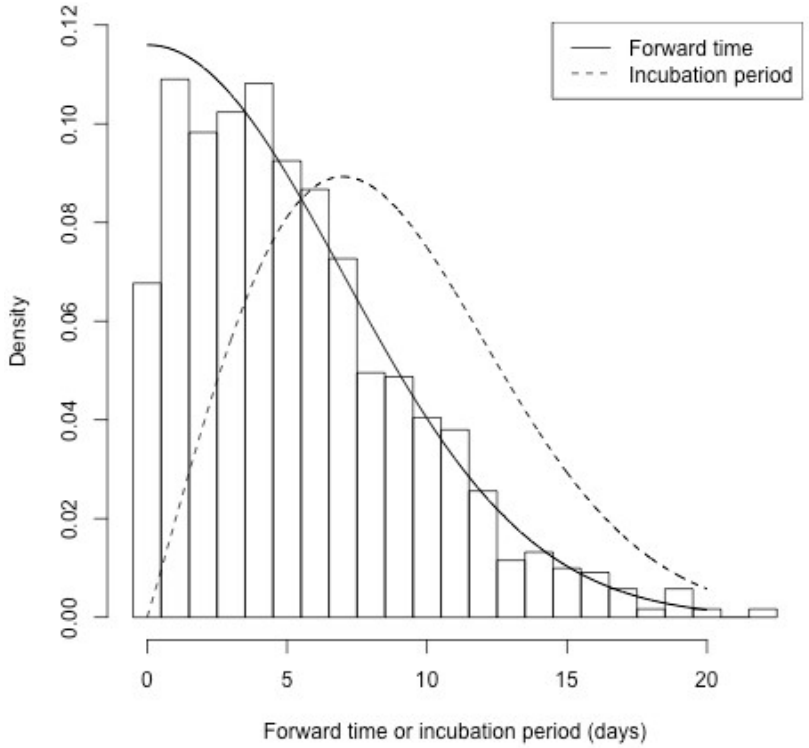

Figure: Histogram and estimated probability density functions for the time from Wuhan departure to symptoms onset, i.e., forward time.

Results

$$\hat{\alpha} = 2.04,\ 95\%\ CI\ (1.80, 2.32)$$

$$\hat{\lambda} = 0.103,\ 95\%\ CI\ (0.10, 0.11)$$

The mean is 8.62 days and median is 8.13 days.
The 90%, 95% and 99% percentiles are, respectively, 14.65, 16.67 and 20.59 days.

Sensitivity

When encompassed in a closed space with large crowds, people are more likely to contract the disease. If one individual contracted disease on the way out of Wuhan, then we observe the incubation period. Otherwise, we observe the forward time.
Let π be the probability that people contracting disease when left Wuhan in airport, train station etc. Then the observed incubation period is a mixture

$$V \sim \pi f(v) + (1-\pi)\frac{\bar{F}(v)}{\mu}$$

For $\pi = 0, 0.05, 0.1, 0.2$, we have conducted a sensitivity analysis.

$$\max_{\alpha,\lambda} \prod_{i=1}^{n} \left[\pi f(v, \alpha, \lambda) + (1 - \pi) \frac{\bar{F}(v, \alpha, \lambda)}{\mu} \right]$$

If $\pi = 0.2$, the $90\%, 95\%$ and 99% percentiles are, respectively, 13.34, 15.37, 19.36 days.

Questions?

Your estimation is based on forward time which is a censored version of incubation period. How do I believe your estimate is real?

Evidence I

Guan et al. (2020) Clinical characteristic of 2019 novel coronavirus infection in China. *N. Engl. J. Med.*
The corresponding author NanShan Zhong, is the leading medical Dr. in China against COVID-19.

Incubation	# cases
14-17 days	13
18-23 days	8
24 days	2

In their published version, only median and inter-quartiles were reported. Why?

1. They treat the 24-day incubation period as an outlier!
2. Perhaps for long incubation period cases, they are not quite sure whether those cases were real!

Meta analysis

We did a meta analysis by matching their median=4 days and inter-quartiles 2 and 7 days with a Weibull distribution. The estimated shape parameter $\alpha = 1.24$ and rate $\lambda = 0.186$. The $90\%, 95\%$ and 99% percentiles are, respectively, 10.54, 13.04 and 18.45 days. If we use the information that two patients (one severe one light case) had 24 days incubation periods, 13 cases (12.7%) had ≥ 14 days, 8 cases (7.3%) had ≥ 18 days, then the percentiles would be longer!

Evidence II

In MedRxiv by Bi...Lessler... (2020). Table S2, reports that the 95% percentile is 14.04 days.
Among travelers, arrival to symptoms onset (among onset after arrival Shen Zheng), the 95% percentile is 13.79 days.
Arrival to symptoms onset is exactly the forward time!

Take home message

We get the largest and cleanest data set based on symptoms onset forward time in the early stage of COVID-19 outbreak. The backward time is not reliable even if available! This is a perfect example to show that probability renewal theory can be used to solve the thorny exposure onset time problem in infectious disease studies. We cannot repeat this study now if we use New York data since COVID-19 cases are everywhere.

In the most extreme case $\pi = 1$ in the mixture model

$$V \sim \pi f(v) + (1 - \pi)\frac{\bar{F}(v)}{\mu},$$

without any modelling, our forward time histogram shows around 5% patients had symptoms onset on or beyond 14 days.

When $\pi = 0$, we have found that 10% patients had symptoms onset on or beyond 14 days.

We conclude around 5% to 10% patients may develop symptom after 14 days!

Acknowledgements

We would like to thank Dr. Dean Follmann at National Institute of Allergy an Infectious Diseases and Professor Mary Thompson at University of Waterloo for many constructive comments which improved this manuscript greatly. We also thank Professor Jing Cheng at UCSF for organizing this Webinar.