All right, and it says the meeting is being recorded.

Okay, so thanks everyone, for coming to this seminar.

And I hope everyone is doing well.

Today, I’m going to talk about some issues of selection bias in early analysis of the COVID-19 pandemic.

You can find the manuscript online, on arXiv, and the slides of this talk is also available on my webpage.

So, here are the three collaborators, involved in this project.

So Nianqiao is a PhD student at Harvard, and we kind of only met online.

We never met in person, and I sort of created a dataset in January, and I wanted some help, somehow she saw this and she said: I could help you.

And we kind of developed a collaboration.

And Sergio and Rajen are both, ah, lecturers in the Stats Lab in Cambridge.

And I’d like to thank many, many people who have given us very helpful suggestions.

This is just some of them.

I’d like to begin with just saying COVID-19 is personal for everyone, and what I would share is partly my story, my personal story with COVID-19.

So here is a photo of me and my parents, taken last September, when I went back to China,
to see my family. So both myself and my parents, we all grow up in Wuhan, China. And on a sunny day in September, we went to, well, this is the Yellow Crane Tower, a sort of landmark building in Wuhan. And the funny thing is, I think I’ve never been there, on top of the tower, in my entire life. This is actually the first time I went there. This is something like if you have a famous local attraction for tourists, you actually don’t go, as a local. And so, on January 23, because the epidemic was growing so fast in Wuhan, it started a lockdown. So, if we went on top of the Yellow Crane Tower, this is what we would see on a typical day, before the lockdown. And on the right, so, there’s sort of what happens after the lockdown, and I liked how the journalist used sort of this gloomy weather as the background, and certainly reflected everybody’s mood, after the lockdown. So, this project begins on January 29. So had a conversation with my parents over the phone, and they told me that a close relative of ours was just diagnosed with, quote/unquote, viral pneumonia.

So, basically at that point, we all think that must
be COVID-19, but because there was not enough tests, this relative could not get confirmed. And this prompted me to start looking through the data available at the time. But I quickly realized that the epidemiological data from Wuhan are very unreliable. And here is some anecdotal evidence. The first evidence is about inadequate testing. So actually this relative of mine could not get an RT-PCR test until mid-February, and she actually developed symptoms on about January 20. So by mid-February, she was already recovering. And she took, I think, several tests. Her first test was actually negative, and a few days later she was tested again, and the result came back positive. So there’s also a lot of false negative tests. I think, in general. And another problem with the epidemiological data from Wuhan is insufficient contact tracing. So, her husband, this relative of mine’s husband, he also showed COVID symptoms, but he quickly recovered from that, and in the end he was never tested for COVID. So, you can also see the insufficient testing from this incidence plot. This is the daily confirmed cases, up until mid-February,
and this is when the travel ban started, or the lockdown started, January 23, and on February 12, there was a huge spike of over 10,000 cases, much more than the previous few weeks.

And the reason for that was not suddenly because people were infected on that date. It’s because of a change of diagnostic criterion. So before February 12, everybody needs to have a positive RT-PCR test to be confirmed a COVID-19 case. But since February 12, because there, the health system in Wuhan was so overwhelmed, the government decided to change diagnostic criterion. So without RT-PCR tests, you can still be diagnosed with COVID-19 if you satisfy several other criteria. And this sort of change in diagnostic criteria only happened in the Hubei Province and not elsewhere in China.

So a solution, if we like to avoid these problems with data from Wuhan, so one clever solution is to use cases that are reported from, sorry, exported from Wuhan. So this has two benefits. First of all, testing and contact tracing were quite intensive in other locations. So, it’s reasonable to expect that a lot of the bias due to sort of under-ascertainment will be less severe.
if we use data from elsewhere. And also, many locations, particularly in some cities in China, published detailed case reports, instead of just case counts. And if you look at these detailed case reports there are a lot of information that can be used for inference. This is not our idea. And I think one of the, at least one of the first persons to use this design was a report from Neil Ferguson’s group in Imperial College, London, and they published a report on January 17, and what it did was a simple sort of division of the number of cases detected internationally, over the number of people traveled from Wuhan, internationally. And they found that it could be over 1,700 cases by January 17, in Wuhan. So, I started this on January 29, and within about two weeks, managed to put something online. Which we also used internationally confirmed cases to estimate epidemic growth. And what we used were 46 coronavirus cases who traveled from Wuhan and then were subsequently confirmed in six Asian countries and regions. And the main result was that the epidemic was doubling in size every 2.9 days.
And we used the Bayesian analysis, and the 95 percent critical interval was two to 4.1.

And of course, when I was writing this article, I was mostly just working on this dataset that we collected, very hard and (muttering), thinking about what model is suitable for this kind of data.

And just before I posted this pre-print, I realized there was a similar article that already published in The Lancet, on January 31.

And what’s really puzzling is they used almost the same data and very similar models, but somehow reached completely different conclusions.

So they used data from December 31 to January 28, that are exported from Wuhan internationally. And they would like to infer the number of infections in Wuhan.

And one of the main results, which was this epidemic doubling time, was 6.4 days, and the 95 percent critical interval was 5.8 to 7.1.

So that’s drastically different from ours.

So again, ours was 2.7, within two to four, and this was 6.4.

And this is talking about the doubling time. So the doubling time of six days versus three days, that’s sort of really, really different.

And the confidence intervals, the credible intervals...
didn’t even overlap. So I was really puzzled by this. And before I tell you what I think, how the Lancet paper got it wrong, I’d like to just show you this plot. You probably have seen this many times before, in news articles, which is just sort of a logarithm of the total cases versus the days, or some time, zero, for each country. And what you see is for both the total number of cases and the total number of deaths, it sort of grew about 100-fold in the first 20 days. At least among these countries that were most hard-hit by COVID-19. And if you just use that as a variable of estimate, of the doubling time, that corresponds to a doubling time of three days. Of course, this is sort of very kind of anecdotal, because this data were not collected in a very careful way, and the amount of cases were not reported, but this is just to show you that perhaps the doubling time of 6.4 days was a bit just, too long. So, towards the end of the talk, I’ll tell you what we think led to these very different results. Just some spoilers, so the crucial difference is that the Lancet study actually did not take into account the travel ban on January 23. And that actually had a very,
very circumstantial selection effect on the data. And this will be made precise later on in the talk. So, for the rest of the talk, I'll first give you an overview of selection bias. Then I'll talk about how we sort of overcome them, by sort of collecting the dataset very carefully and building a model very carefully. And then I'll talk about why the Lancet study I just mentioned and some other early analysis were severely biased. If there is time, I will tell you a little bit about our Bayesian nonparametric model. And then I'll give you some lessons I learned from this work. So selection bias. So we identified at least five kinds of selection bias in COVID-19 studies. So the first one is due to under-ascertainment. So this may occur if symptomatic patients do not seek healthcare, or could not be diagnosed. So this kind of bias, and basically what we can do is to,
to think about a clever design to avoid this problem,
to focus on locations where the testing is intensive.
The second bias is due to non-random sample selection.
So, basically this means that the cases included in the study are not representative of the population.
So this essentially applies to all studies, because detailed information about COVID-19 cases are usually sparse; they’re not always published.
But especially for studies that do not have a clear inclusion criterion, and if they just sort of simply collect data out of convenience, then there could be a lot of non-random sample selection bias.
And again, statistical models are not really gonna help you with this kind of bias.
You’d use, you’d follow some protocol for data collection, and you would exclude some data that do not meet the sample inclusion criterion.
Even when that may, leads to inefficient estimates.
The third bias is due to the travel ban.
This is kind of my spoiler about that Lancet study.
So basically, outbound travel from Wuhan to anywhere else was banned from January 23 to April eight.
So if the study analyzed cases exported from Wuhan, then they’re susceptible to this selection defect.
And this would usually lead to underestimation of epidemic growth, and the reason is that, so, the epidemic is growing very fast, but then you essentially can’t observe cases that were supposed to leave Wuhan after January 23. So if you just wait for a long time, and then look at the epidemic curve among the cases exported from Wuhan, it may appear that, ah, it sort of dies down a little bit, but that’s not because of the epidemic being controlled. That’s because of the travel ban. Fortunately this bias, you can correct for it by deriving some likelihood function tailored for the travel restrictions. The fourth bias is ignoring the epidemic growth, and basically if you think about people who have been in Wuhan before January 23, they’re much more likely to be infected towards the end of their exposure period than early, and that’s because the epidemic was growing quickly. So, there are many studies, or I should say there are several studies of the incubation period that simply treat infections as uniformly distributed over the patients’ exposure period to Wuhan. And this will lead to overestimation of the incubation period.
Because actually, the infection time is much, much closer to sort of the end of their exposure.
And this is also a bias that can be corrected for, by doing statistical analysis carefully.
The fifth and last bias is due to right-truncation. So this happens in early analysis because,
and this could lead to some right-truncation bias. And this generally would lead to underestimation
of the incubation period. So this is, so incubation period, I forgot to mention,
is just the time between infection to showing symptoms.
So, right-truncation would lead to underestimation
of incubation period, because people with longer
incubation period may not have showed symptoms
by the time that these datasets were collected.
So the solution to this is we need to both collect cases
that meet the selection criterion, and continue
data collection until a sufficiently long time.
Or, you derive some likelihood function to correct
for the right-truncation.
So we’ll go over this later.
So just to recap, on a very high level, there are at least five kinds of biases in COVID-19 analysis. And if you read sort of article pre-prints or use articles, I think you will find some kind of resemblance of these biases in many studies. The keys to avoid selection bias is basically, you just do everything carefully. You design the study carefully, collect the sample carefully, and analyze the data carefully. But the reality, of course, is not that simple. And what I will show below, it’s an example of our try to eliminate or to reduce selection bias as much as possible.

So, let me tell you the dataset we collected. We found 14 locations in Asia, some are international, so Japan, South Korea, Taiwan, Hong Kong, Macau, Singapore. Some are sort of in mainland China. So there are several cities in mainland China. So all these locations have published detailed case reports from their first local case. So, most of the Chinese locations, I mean, they were done with the first wave of the epidemic by the end of February. So Japan, Korea and Singapore saw some resurgence.
of the epidemic later on, and eventually, they did not publish detailed case reports. But for our purposes, these locations all published detailed reports before mid-February, and that’s about three weeks after the lockdown of Wuhan.

So it’s pretty much enough to find out all the Wuhan exported cases. So just to give you a sense of the kind of data we collected, this is sort of all the important columns in the dataset, and the particularly important columns are marked in red.

So, we collected, there was a case ID, where the case lived, the gender, the age, whether they had known epidemiological contact with other confirmed cases, whether it has known relationship with other confirmed cases. So most of the time, this is relatively easy to fill. For example, if you’ve never been to Wuhan, this entry must be yes. But sometimes, this can be a little bit tricky.

So this column, outside column means that, whether we think the data collector thinks this case is transmitted outside Wuhan. Most of the time, this is relatively easy to fill. For example, if you’ve never been to Wuhan, this entry must be yes.
For example, this person, the fifth case in Hong Kong, is the husband of the fourth case in Hong Kong, and they traveled together from Wuhan to Hong Kong. So it’s unclear if this case is transmitted in or outside Wuhan, so we put a “likely” there. And the other information are some dates, the beginning of stay in Wuhan, the end of stay in Wuhan, the period of exposure, which would equal to beginning to the end of stay in Wuhan, for Wuhan exported cases, but can be different for other cases. When the person showed symptoms, when did they first go to a hospital, and when were they confirmed a COVID-19 case. When they are confirmed a COVID-19 case. When the person showed symptoms. When the case arrived at a final location where they are confirmed a COVID-19 case. When did they first go to a hospital, and when were they confirmed a COVID-19 case. So we collected about 1,400 cases with all this information. And overall, I think our dataset was relatively high in quality, and most of the cases had known symptom onset dates; only nine percent of them have that entry missing.

So, one important step after this is to find out which cases are actually exported from Wuhan. So I’ve been using this terminology from the beginning.
of the talk, but basically the case is Wuhan exported.

if they are infected, if they were infected in Wuhan.

And then confirmed elsewhere.

So we had a sample selection criterion.

to discern a Wuhan exported case.

I’m not going to go over it in detail,

but basically the principle we followed.

is that we would only consider a case as Wuhan exported.

if it passed a beyond a reasonable doubt test.

So basically, if we think there is a reasonable doubt.

that the case could be infected elsewhere,

then we would say: let’s exclude that from the dataset.

So this eventually gives us 378 cases.

Next I’m gonna talk about the model we used.

So the model is called: BETS.

It’s named after sort of four key epidemiological events.

The beginning of exposure, the end of exposure,

time of transmission, which is usually unobserved,

and the time of symptom onset, S.

So what we will do below is we’ll first define the support.

of these variables, so we call that P.

Which is basically represents the Wuhan exposed population.

So this is the population we would like to study.

We will then construct a generative model.

for these random variables.

Basically, for everyone in the Wuhan exposed population.
Then, to consider the sample selection, we'll define a sample selection set, \( D \), that corresponds to cases that are exported from Wuhan. Then finally we will derive likelihood functions to adjust for the sample selection. So essentially, what we're trying to infer is the disease dynamics in the population, \( P \), but we only have data from this sample, \( D \). So here's a lot of work that needs to be done to correct for that sample selection. So intuitively, this population \( P \) are just all people who have stayed in Wuhan, between December first and January 24, so anyone who has been in Wuhan for maybe even just a few hours, they would count as someone exposed to Wuhan. And I'm going to make some conventions to simplify this set, \( P \), a little bit. So \( B = 0 \) has a special meaning. So \( B = 0 \) means that they actually started their stay in Wuhan before time zero, so they live in Wuhan essentially. So \( B > 0 \) means these other cases visited Wuhan sometime in the middle of this period, and then they left Wuhan. So \( E = \infty \) means that the case did not arrive in the 14 locations we are considering.
before this lockdown time, \( L \).

So for the purpose of our study, we did not need to differentiate between people who have always stayed in Wuhan past time \( L \), or people who left Wuhan before time \( L \), but went to a different location.

So \( T = \infty \) means that the cases were not infected during their stay in Wuhan.

So this could be infected outside Wuhan, or it could be they were never infected.

And \( S = \infty \) means that the case did not show symptoms of COVID-19, and it can simply be, they were never infected. Or the case was actually tested positive for COVID-19, 

but never showed symptoms, so it’s, they’re asymptomatic.

So under these conventions, this is the set, this is the support for this population, \( P \).

So \( B \) is between zero and \( L \), \( E \) is between \( B \) and \( L \) or \( \infty \), \( T \) is between \( B \) and \( E \), which means that they are, in fact, in Wuhan, or \( \infty \).

And \( S \) is between \( T \) and \( \infty \), and \( S \) can be equal to \( \infty \).

So now we have defined this population, \( P \).

And now let’s look at a general model, a data-generated model for this population.

So, by the basic law of probability, we could decompose the joint distribution.
00:31:21.160 --> 00:31:25.460 of BETS into these four, and the first two
00:31:25.460 --> 00:31:27.210 are the distribution of B and E.
00:31:27.210 --> 00:31:29.520 They are related to travel.
00:31:29.520 --> 00:31:32.370 The second one, sorry, the third one is the distribution
00:31:32.370 --> 00:31:34.790 of T given B and E.
00:31:34.790 --> 00:31:37.920 So that’s about the disease transmission.
00:31:37.920 --> 00:31:40.600 And the last one is the distribution of S,
00:31:40.600 --> 00:31:44.583 given BET, and that’s related to disease progression.
00:31:46.900 --> 00:31:49.610 So we need to make two basic assumptions,
00:31:49.610 --> 00:31:54.490 and they are important because we would like to infer
00:31:54.490 --> 00:31:56.500 what’s going on in the population P,
00:31:56.500 --> 00:32:01.500 from the sample T, from these Wuhan exported cases.
00:32:01.960 --> 00:32:05.290 So we need to sort of make assumptions
00:32:05.290 --> 00:32:08.320 so we can make that extrapolation.
00:32:08.320 --> 00:32:10.510 So the first assumption, we assume it’s about
00:32:10.510 --> 00:32:14.100 this disease transmission, and it basically means
00:32:14.100 --> 00:32:17.163 that the disease transmission is independent of travel.
00:32:18.350 --> 00:32:22.490 So there is a basic sort of function that’s independent
00:32:22.490 --> 00:32:25.803 of the travel that’s growing over time.
00:32:27.000 --> 00:32:31.293 And then there’s the rest of the points mass at infinity.
00:32:32.790 --> 00:32:36.840 This T function, so, it will appear later on.
00:32:36.840 --> 00:32:38.883 It’s the epidemic growth function.
00:32:40.160 --> 00:32:43.240 The second assumption is that the disease progression
00:32:43.240 --> 00:32:45.203 is also independent of travel.
00:32:46.420 --> 00:32:49.470 So, what’s assumed here is basically
that there is one minus $\mu$ of the infections, that are asymptomatic in that they didn’t show symptoms.

The amount of people who showed symptoms, the incubation period, which is just $S - T$, follows this distribution, $H$.

Okay, so $H$ is the density of the incubation period, for symptomatic cases.

And this whole distribution does not depend on $B$ and $E$.

So these are sort of the two basic assumptions that we relied on.

There are two further parametric assumptions that were useful to simplify the interpretation, but they can be relaxed.

So the next, one assumption is the epidemic was growing exponentially before the lockdown.

And then that, the other assumption is that the incubation period is gamma-distributed, okay?

So there’s some parameters, kappa, $R$ and alpha, beta.

So, don’t worry about nuisance parameter $\mu$, which is the proportion of asymptomatic cases. And kappa, which is some baseline transmission.

So it turns out that they would be canceled in the likelihood function, so they won’t appear in the likelihood function.

And (muttering) these parametric assumptions, they can be relaxed and they will be relaxed in the Bayesian parametric analysis, if I can get to there.
But essentially, these are very useful assumptions that allow us to derive formulas explicitly. So I have covered the full data BETS model for the population P.

Now we need to look at what we can observe. So what we can observe are people in B that satisfy three additional restrictions. The first restriction is that the transmission is between their exposure to Wuhan. The second restriction is that the case needs to leave Wuhan before the lockdown time, L. The third restriction is that the case needs to show symptoms. So S is less than infinity.

So some of the locations we considered did report a few asymptomatic cases, but overall, asymptomatic ascertainment was very inconsistent. So we only considered cases who showed symptoms.

So this gives us the set of samples that we can observe in our data. So, which likelihood function should we use? For a moment, let’s just pretend that the time of transmission, T, is observed.

So if we had samples, ID samples from the population, P, then we could just use this product of the density of BETS as a likelihood function. But this is not something we should use, because we actually don’t have samples from P.
We have samples from $D$, so what we should do is to condition on the selection set, $D$, and use this likelihood function, which is basically just the density divided by the probability that someone is selected in this set, $D$. Okay, this is called unconditional likelihood, to contrast with the conditional likelihood. So, in unconditional likelihood, we consider the joint distribution of $B$, $E$, $T$, and $S$. But in the conditional likelihood, we consider the conditional distribution of $T$ and $S$, given $B$ and $E$. So this is the conditional distribution of the disease transmission and progression, given the travel. So this treats travel as fixed. So to compute this conditional likelihood, we need further conditions on $B$ and $E$, okay? But in reality, the time of transmission, $T$, is unobserved, so we cannot directly use the likelihood function, as on the last slide, so one possibility is to treat $T$ as a latent variable and use, for example, an EM algorithm. The way we chose is to use an integrated likelihood. That just sort of marginalized over this unobserved variable, $T$. So, the unconditional likelihood is the product over the cases of the integral.
00:38:26.250 --> 00:38:29.733 of the density function over T.
00:38:31.070 --> 00:38:34.380 And the conditional likelihood is just a product
00:38:34.380 --> 00:38:39.193 of the integral of the conditional distribution of T
00:38:40.210 --> 00:38:41.043 and S,
00:38:44.750 --> 00:38:48.690 So, the reason we sort of considered both
00:38:48.690 --> 00:38:51.160 the unconditional likelihood and conditional like-
00:38:51.160 --> 00:38:55.050 lihood
00:38:55.050 --> 00:39:00.000 is that the unconditional likelihood is a little bit
00:39:00.000 --> 00:39:05.000 more efficient, because it also uses information
00:39:05.840 --> 00:39:08.090 in this density, BE, given your selected.
00:39:09.228 --> 00:39:12.040 So that contains a little bit of information.
00:39:12.040 --> 00:39:17.040 But a conditional likelihood is more robust.
00:39:17.940 --> 00:39:22.163 So, because it does not need to specify how people
00:39:24.000 --> 00:39:28.883 traveled,
00:39:28.883 --> 00:39:32.338 so it is robust to misspecifying those distributions.
00:39:35.620 --> 00:39:37.283 Is this clear to everyone?
00:39:39.570 --> 00:39:41.663 If so, I’m gonna proceed.
00:39:44.850 --> 00:39:49.010 Okay, so under these four assumptions
00:39:49.010 --> 00:39:52.580 that I introduced earlier, you can sort of compute
00:39:52.580 --> 00:39:56.680 the explicit forms of the conditional likelihood
00:39:56.680 --> 00:39:59.390 I’m not gonna go over the detailed forms,
00:39:59.390 --> 00:40:01.940 but I just want to point out that first of all,
00:40:01.940 --> 00:40:04.420 as I mentioned earlier, this does not depend on
00:40:04.420 --> 00:40:07.203 the two nuisance parameters, mu and kappa.
00:40:08.190 --> 00:40:12.380 And second of all, this actually reduces to a like-
00:40:12.380 --> 00:40:17.380 lihood
00:40:17.380 --> 00:40:18.970 function that’s previously derived in this paper in
00:40:18.970 --> 00:40:21.940 2009
00:40:21.940 --> 00:40:24.010 So R equals to zero means that the epidemic
was not growing, so it’s mostly a stationary epidemic.

So that’s reasonable for maybe influenza, but not for COVID.

So for unconditional likelihood, we need to make further assumptions about how people traveled, the assumption we used was just a very simple, sort of a uniform assumption, that assumes that the travel was stable in the period that we considered.

And we use those assumptions, we can derive the closed form unconditional likelihood.

There’s a little bit of approximation that’s needed, but that’s very, very reasonable in this case.

So, I’d like to show you the results that fit in these parametric models.

So what we did is we obtained point estimates of the parameters by maximizing the likelihood functions I just showed you, and then we obtained 95 percent confidence intervals, by a likelihood ratio test. So, what you can see is broadly, over different locations, the estimated doubling time was very consistent.

Also cross-conditional and unconditional likelihood,

so the doubling time was about two to 2.5 days. And the median incubation period is about four days,

but there is a little bit of variability in the estimates.
It turns out that the variability is mostly because of the parametric assumptions that we used. And then the 95 percent quantile is about, 12 to 14 days. Or if you consider the sampling variability, that is about 11 to 15 days. Okay, but broadly speaking, across the different locations, they seem to suggest very similar answers. So, just to summarize, the initial doubling time seems to be between two to 2.5 days. Median incubation period is about four days, and 95 percent quantile is about 11 to 15 days. So, those sort of were our results, using the parametric model. And next I’m going to compare it with some other earlier analysis, and give you a demonstration, or an argument of why some of the other early analysis were severely biased. So first, let’s look at this Lancet paper that I mentioned in the beginning of the talk that estimated doubling time. So the doubling time they estimated was 6.4. days. So, what happened is these authors used a modified SEIR model, so the SEIR model is very common in epidemic modeling, so the modified that model to account for traveling, but they did not account for the travel ban. So, basically to sort of simplify what’s going on,
what they essentially did is they used the density of the symptoms as in the population P, so they fitted this density, but they fit it using, ah, samples from the set D. So it is quite reasonable to assume that the incidence of symptom onset was growing exponentially in the population that is exposed to Wuhan. So given P, this distribution, margin distribution of S, was perhaps growing exponentially before the lockdown. But we don’t actually have samples from P. We have a sample from D. So, we actually can derive the density of S and D, and that looked very different from exponential growth. So, basically the intuition is that if you look at the distribution of the transmission, T, it is growing exponentially, but it also has this effect, this exponential RT times L minus T. So basically, if you are transmitted on time T, then you only have the time between T to L to leave Wuhan and be observed by us. Okay, so that’s why it’s not only exponential growth, but there’s also a decreasing trend, L minus T, for the distribution of the time of transmission.
convolved with the distribution of the incubation period.
And that has this form that is approximately an exponential growth, and then times this term, that is \( L \) plus some quantity that depends on the incubation period and the epidemic growth, minus \( S \).
So this is a term that is not considered, in this simple exponential growth model. Which is basically what’s used in that Lancet paper.
Okay, so to illustrate this, what I’m showing you here is a histogram of the symptom onset of all the Wuhan exported cases, who are also residents of Wuhan.
So they stayed from December first to January 23. What you see is that it was kind of growing very fast, perhaps exponentially in the beginning, but then it slows down around the time of the lockdown.
Okay, so the orange curve is the theoretical fit that we obtained in the last slide, using the maximum likelihood estimator of the parameters. So it fits the data quite will.
So what happened, I think, with the Lancet paper is, basically stopped about January 28th, so it’s about here, and they essentially tried to fit an exponential growth from the beginning to January 28.
And that would lead to much faster growth than fitting the whole model to account for the selection.

Okay.

So that’s about epidemic growth.

Next I will talk about several studies of the incubation period.

These studies are susceptible to two kinds of biases.

One is that some of them ignore the epidemic growth, so instead of using this likelihood function, this conditional likelihood function, to just fit this $R$ is equal to zero, and then they use this likelihood function that was derived in the early paper.

The other bias is sort of right-truncation. And basically, they kind of stopped the data collection early and only used cases confirmed by then, so people with long incubation periods are less likely to be included in the data, so that leads to underestimation of the incubation period.

And a solution to this is you can actually derive the likelihood with additional conditioning events, that $S$ is equal, sorry, less than or equal to some threshold, $M$.

Suppose you stop the data collection a week after $M$, and you say: perhaps we have all, find out all the cases who showed symptoms beforehand.
We can use this likelihood function. I’m not gonna show you the exact form, but basically you need to further divide by, ah, the probability of S less than or equal to M, and you can obtain closed-form expression for this under our parametric assumptions.

Using integration by parts.

So, I’d like to show you an experiment to illustrate this selection bias.

So in this experiment, we kind of stop the data collection between any day from January 23 to February 18, and we fitted sort of this parametric BETS model, using one of the following likelihood.

So this is the likelihood that treats R equals to zero, so it’s adjusted for nothing,

and this is the likelihood derived earlier and used in other studies.

This is the likelihood function that adjusts for the growth, so R is treated as an unknown parameter.

And this is the likelihood on the last slide that adjusted for both the growth and the right-truncation, as less than or equal to M.

So the point estimates are obtained by MLEs, and the confidence intervals are obtained by nonparametric Bootstrap, and we compared our results with three previous studies.

So this is, basically summarizes this experiment.
This is a little bit complicated, so let me walk you through slowly. There are three likelihood functions we used. One adjusts for nothing; that’s the orange. The one is adjusted only for growth, and the ones that adjusted for both growth and truncation. Okay, so what you can immediately see is that if we adjusted for nothing, and if you sort of used our entire data set, the median incubation period would be about nine days. And the 95 percent quantile would be about 25 days. So that’s just way too large. And if you ignored right-truncation, for example, that only accounts for growth, you underestimate the incubation period in the beginning, as expected, but you slowly converge to this final estimate. And if you use this likelihood function and adjust for both growth and truncation, you actually get some quite sensible results by the end of January. So, it has large uncertainty, but it’s roughly unbiased, and it kind of eventually converges to that estimate.
The same estimate that we obtained using the blue curve, but using the full data. Okay.

So, for the sake of time, I think I’ll skip the part about Bayesian nonparametric inference.

One thing that’s a little bit interesting, I think, is there seems to be some difference between men and women in their incubation period.

So these are sort of the posterior mean and posterior credible intervals for nonparametric incubation period, and you can see that men seem to develop symptoms quicker than women.

So, that’s a little bit interesting, and maybe, I mean, I’m not a doctor, but it could be related to the observation that men seem to be more susceptible and die more frequently than women.

So let’s, let me conclude this talk.

So these are some conclusions we found about COVID-19.

Initial doubling time in Wuhan was about two to 2.5 days.

The median incubation period is about four days, and the proportion of incubation period above 14 days is about five percent.

There are a number of limitations for our study. For example, we used the symptom onset reported by the patients and they are not always accurate.

There could be behavioral reasons for people to report a later symptom onset.
Even though these locations are intensive in their testing
and contact tracing, some degree of under-ascertainment
is perhaps inevitable.
As I have shown you, in our dataset collection,
discerning the Wuhan exported case is not a black and white decision.
We used this beyond a reasonable doubt kind of criterion,
but that’s one criterion you can apply.
And the crucial assumptions, we put the first two assumptions, which means that the travel and disease are independent, and that can be violated.
For example, if people tend to cancel their travel plans when feeling sick.
Nevertheless, I think I have demonstrated some very compelling evidence for selection bias in early studies.
Some of the biases you can correct by designing the study more carefully, some require more sophisticated statistical adjustments.
And basically, I think the conclusion is:
you should make un-calculated BETS.
So, we should always carefully design the study and adhere to our sample inclusion criteria.
And the statistical inference should not be based on some intuitive calculations,
but should be based on first principles.
In this study, we kind of went back all the way
to defining the support of random variables. So that’s sort of statistics 101. But that’s actually, it’s extremely important. So I found it really helpful to start all the way from the beginning and develop a generative model. And that avoids a lot of potential selection biases. So the final lesson I’d like to share from this whole study is that I think this demonstrates the data quality and better design are much more important than data quantity and better modeling, in many real data studies. Thanks for the attention, and I’ll take any questions from here. - Thanks to you for the nice talk. Does anyone have questions for Qingyuan? - Okay. - Are there any information in datasets of whether patient is healthcare worker? - No, these are not usually healthcare workers. These are exported from Wuhan, so they’re usually just people who traveled maybe for sightseeing, or for the Chinese New Year, they traveled from Wuhan to other places and were diagnosed there. - Right, so also he has another question. - Joe has another question also: how can we evaluate
the effectiveness of social distancing and mask guidelines?

- I think this study we did was not designed to answer those questions.

We did have a very, ah, sort of preliminary analysis. So we broke the study period into two parts. On January 20, it was confirmed publicly that the disease was human-to-human transmissible, so we broke the period into two parts: those before January 20 and those after January 20.

But the after period is just three days. So January 21, 22, 23, and we found that if we fit different growths to these two periods, the second period, it seemed that the growth was substantially slower. The growth, the exponent R is not quite zero, but it’s close.

So it seems that the knowledge of sort of human-to-human transmissibility and the fact that, I think, masks are probably much more, were much more available in Wuhan, people started to do some social distancing right after January 20.

I think that seemed to play a role. But that’s very, very preliminary, and I think there are a lot of good studies about this now.

- Donna has a question.
00:59:26.400 --> 00:59:31.400 Donna, do you want to say what your question is?
00:59:32.180 --> 00:59:33.140 - [Donna] Yeah, sure, thanks.
00:59:33.140 --> 00:59:36.230 That was a very interesting and clear talk.
00:59:36.230 --> 00:59:39.850 I really appreciated the way you carefully went through,
00:59:39.850 --> 00:59:43.907 step by step, to show-- (audio distorting)
00:59:47.449 --> 00:59:49.650 Who aren’t doing that, I feel.
00:59:49.650 --> 00:59:53.770 But my question was, it was still hard for me to tell
00:59:53.770 --> 00:59:58.770 to what extent your estimates were identifiable
00:59:59.420 --> 01:00:04.270 due to assumptions and to what extent the data
01:00:04.270 --> 01:00:07.293 made the estimates fairly identifiable.
01:00:08.640 --> 01:00:11.533 - Yeah so essentially, I mean, selection bias,
01:00:12.430 --> 01:00:17.040 usually you cannot always avoid it, unless you
01:00:17.040 --> 01:00:22.000 make some kind of missing at random type of assumption.
01:00:22.000 --> 01:00:24.650 Here, we don’t have a random selection.
01:00:24.650 --> 01:00:26.950 It’s more like a deterministic selection,
01:00:26.950 --> 01:00:30.060 and we can quantify that selection event,
01:00:30.060 --> 01:00:35.060 but still, as you said, I think these are great ques-
01:00:36.590 --> 01:00:41.321 to sort of disentangle the nonparametric assump-
01:00:41.321 --> 01:00:44.246 tions
01:00:44.246 --> 01:00:45.463 needed for identification and the parametric ass-
01:00:50.600 --> 01:00:52.740 I don’t have a formal result,
01:00:52.740 --> 01:00:56.350 but my feeling is the first two assumptions
01:00:56.350 --> 01:00:59.920 that are assumed, sort of the independence of travel
01:00:59.920 --> 01:01:04.920 and disease, that’s sort of essential to the identi-
34
And then later on, the assumptions are perhaps relaxable. So we did try to relax those in the Bayesian nonparametric analysis. But that’s not a proof, so that’s my, ah, best guess at this point.

From Casey, said, ah, the estimates, people have estimated about five to 80 percent of asymptomatic infections, and isn’t that a limitation of your model that you did not account for asymptomatic carriers? And if so, how can we possibly model for it, given the large range of estimates? So this is actually a feature of our study, because we actually had a, let’s see, we had a term for the asymptomatic transmission. So, but that’s just that parameter was canceled. So this parameter, mu, or one minus mu, is the proportion of asymptomatic infections. But then because we only observed cases who are, who showed symptoms, so actually in likelihood, this parameter mu got canceled. Of course the reason we could cancel that mu is because of this assumption, too, that S is independent of the travel. So that’s important. But once you assume that you actually, ah, sort of don’t need to worry about asymptomatic transmission,
and on the other hand, this dataset, or this whole method also provides more information about the proportion of asymptomatic infection. Hopefully that’ll answer your question. - [Casey] Yeah, thanks; so you account for it by saying it’s not really significant, in your estimate? - Yeah, so in the likelihood, you will get canceled. So it doesn’t appear in the likelihood. So the likelihood of the data does not depend on how much are asymptomatic, because we only look at cases who are symptomatic. So this incubation period that we estimated are also the incubation period among those people who showed symptoms. So it’s an elegant way of sidestepping the question, (laughing) in a way. Well, it’s not a sidestep, it’s sort of, it’s a limitation of this design. So the whole design should be robust to asymptomatic transmission, and it also gives no information about asymptomatic transmission. Yeah, I was really impressed at the way you took on that Lancet article and just really, ah, it was really impressive; what a great talk. Thank you so much. Well thank you. Hi Qing I have a question.
So you mentioned before that because the measurements inside of Wuhan are the, or the, ah, the measurements that we have inside Wuhan, the numbers aren’t very accurate due to various reasons. So I’m wondering that if you calculate the doubling time using the data for Wuhan city, and then take into, that uses the measurements before they changed the criterion for when it’s counted before that date, will you get a similar measurement, a similar estimate as if you’re using the traveling data, or is it much worse?

What I would like to point out is that this figure is only the number of new, confirmed cases. So what is usually done in epidemic analysis is they don’t look at the number of confirmed cases, but the number of cases who showed symptoms on a certain day because that’s usually less variable, less noisy, than this sort of confirmation.
So people have done that, and I don’t see a doubling time estimation from that; there was a journal paper on that.

And there was also a very interesting comment on it that criticized some of its methodology. I didn’t see a doubling time estimate. So they seemed to focus on the R-naught of the epidemic.

I actually had thought about that as well, and we, in this study I have presented, intentionally avoided to estimate R-naught. Because I think there was a lot of issues with, ah, finding out the unbiased estimate of the serial interval, which is very important in estimating R-naught. So, this estimate we found is not directly comparable to that journal paper, I guess.

But so what happened, I think, is around late January, early February, all of people have tried to estimate the R-naught and the doubling time of the epidemic, and what I’ve found interesting was there were kind of two modes.

There’s several papers estimated that the doubling time was about six to seven days, and there were several papers that estimated doubling times of about two to four days.
at least I have shown that the Lancet paper, that their whole method seems to be very flawed. But whether this means that our estimate is very close to the truth, it doesn’t necessarily mean so. Because we also have a lot of limitations.

Any more question for Qingyuan? Okay, thanks. I guess that’s all for today, and it’s a great talk. If you have any more questions for Qing, you can send him an email, and you can find his email on his website, okay?

Okay. All right, okay, thank you everyone. Thank you, oh, we got a new message? It’s just a, Keyong said thank you. Okay, okay, bye! [Qingyuan] All right, bye.