学 位 論 文

Estimation and application of epidemiological parameters of COVID-19

(新型コロナウイルス感染症における疫学パラメータの推定と応用)

2021年 12月

北海道大学

Natalie Marie Linton

（ナタリー・リントン）

学 位 論 文

Estimation and application of epidemiological parameters of COVID-19

(新型コロナウイルス感染症における疫学パラメータの推定と応用)

2021年 12月

北海道大学

Natalie Marie Linton

（ナタリー・リントン）

Table of contents

List of publications and presentations 1

Summary 3

List of abbreviations 6

Introduction 7

Chapter 1: Early estimates of the incubation period and other epidemiological parameters of COVID-19 10

Chapter 2: Localized end-of-outbreak determination for coronavirus disease 2019 (COVID-19): examples from clusters in Japan 30

Chapter 3: Correlation between times to SARS-CoV-2 symptom onset and secondary transmission undermines epidemic control efforts 52

Conclusion 94

Acknowledgements 97

Disclosure of conflicts of interest 98

References 99

List of publications and presentations

Publications in this thesis:

1. Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, Jung S-m, Yuan B, Kinoshita R, Nishiura H. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *J Clin Med*. 2020;9(2)538.
2. Linton NM, Akhmetzhanov AR, Nishiura H. Localized end-of-outbreak determination for coronavirus disease 2019 (COVID-19): examples from clusters in Japan. *Int J Infect Dis*. 2021;105:286-292.
3. Linton NM, Akhmetzhanov AR, Nishiura H. Correlation between times to SARS-CoV-2 symptom onset and secondary transmission undermines epidemic control efforts. Under review.

Select related publications:

4. Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. *Int J Infect Dis*. 2020;93:284-286.
5. Nishiura H, Linton NM, Akhmetzhanov AR. Initial cluster of novel coronavirus (SARS-CoV-2) infections in Wuhan, China is consistent with substantial human-to-human transmission. *J Clin Med*. 2020;9(2):488.
6. Nishiura H, Kobayashi T, Miyama T, Suzuki A, Jung S-m, Hayashi K, Kinoshita R, Yang Y, Yuan B, Akhmetzhanov AR, Linton NM. Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). *Int J Infect Dis*. 2020;94:154–155.
7. Anzai A, Kobayashi T, Linton NM, Kinoshita R, Hayashi K, Suzuki A, Yang Y, Jung S-m, Miyama T, Akhmetzhanov AR, Nishiura H. Assessing the impact of reduced travel on exportation dynamics of novel coronavirus infection (COVID-19). *J Clin Med*. 2020;9(601).
8. Jung S-m, Kinoshita R, Thompson R, Linton NM, Yang Y, Akhmetzhanov AR, Nishiura H. Epidemiological identification of a novel pathogen in real time: Analysis of the atypical pneumonia outbreak in Wuhan, China, 2019–2020. *J Clin Med* 2020;9(3):637.

Research presented in this thesis has been shared at conferences as listed below:

1. “Correlation between times to SARS-CoV-2 symptom onset and secondary transmission undermines epidemic control efforts.” 5th Workshop on Viral Dynamics. October 4–5, 2021. Poster presentation. Online. International conference.

2. “Variation in the serial interval distribution among reported cases in Japan.” Society for Mathematical Biology 2021. June 16, 2021. Oral presentation. Online. International conference.

3. “First and second wave serial interval of COVID-19 in Japan.” The 91st Annual Meeting of the Japanese Society for Hygiene. March 8, 2021. Oral presentation. Online. Domestic conference.

4. “Localized end-of-outbreak determination for COVID-19: Examples from clusters in Japan.” The 31st Annual Scientific Meeting of the Japanese Epidemiological Association. January 28, 2021. Oral presentation. Online. Domestic conference.

5. “Localized end-of-outbreak determination for COVID-19: Examples from clusters in Japan.” COVID-19 Dynamics & Evolution Virtual Conference Series. October 19, 2020. Oral presentation. Online. International conference.

Summary

Background and Objectives: Coronavirus disease 2019 (COVID-19) emerged as a human-to-human transmissible disease in late 2019, with its causal pathogen identified as severe acute respiratory disease coronavirus 2 (SARS-CoV-2) in January 2020. Identification of key characteristics of the natural history of SARS-CoV-2 infections was necessary for understanding transmission patterns of the newly emerged disease, as well as for investigating the likelihood that transmission would die out, as had happened with another coronavirus—severe acute respiratory disease (SARS)—in 2005. Chapter 1 estimates the incubation period—the time between infection and illness onset—of COVID-19 cases reported in January–February 2020 to establish a recommended quarantine period. Time from illness onset to hospital admission, time from illness onset to death, and time from hospital admission to death—parameters which are useful for estimating understanding hospitalization and fatality rates—were also estimated. Next, Chapter 2 explores end-of-outbreak probabilities for clusters of COVID-19 cases Japan, which adopted a cluster-based approach to COVID-19 response from February 2020. Lastly, Chapter 3 identifies transmission pairs in Japan and determines the correlation between joint estimates of the incubation period and generation interval—the time between infection of an individual with a pathogen and transmission of that pathogen to another individual—of COVID-19 to evaluate the effectiveness of case isolation on controlling pathogen transmission.

Chapter 1: Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data

Methods: The geographic spread of SARS-CoV-2 infections from the original epicenter of Wuhan, China, in early 2020 provided an opportunity to study the natural history of the recently emerged viral disease. Using publicly available event-date data from COVID-19 cases reported in January 2020 we investigated the incubation period and other time intervals that govern the epidemiological dynamics of SARS-CoV-2 infections. **Results:** The incubation period fell within the range of 2–14 days with 95% confidence and had a mean of around 5 days when approximated using the best-fit lognormal distribution. The mean time from illness onset to hospital admission (for treatment and/or isolation) was estimated at 3–4 days without truncation and at 5–9 days when right-truncated. **Discussion:** Based on the 95th percentile estimate of the

incubation period, we recommended that the length of quarantine should be at least 14 days. The median time delay of 13 days from illness onset to death (17 days with right truncation) should be considered when estimating the COVID-19 case fatality risk.

Chapter 2: Localized end-of-outbreak determination for coronavirus disease 2019 (COVID-19): examples from clusters in Japan

Methods: End-of-outbreak declarations are an important component of outbreak response as they indicate that public health and social interventions may be relaxed or lapsed. We assessed end-of-outbreak probabilities for clusters of COVID-19 cases detected during the first wave of the pandemic in Japan. We computed a statistical model for end-of-outbreak determination using the offspring distribution—the distribution of the number of secondary cases infected by each case—and serial interval—the time from illness onset in a given infected case to illness onset in a case they infect—and accounted for the reporting delay for new cases. Four clusters representing different social contexts and time points during the first wave of the epidemic were selected and their end-of-outbreak probabilities were evaluated. **Results:** The speed of end-of-outbreak determination was most closely tied to outbreak size. Notably, accounting for underascertainment of cases led to prolonged end-of-outbreak determinations. In addition, end-of-outbreak determination was closely related to estimates of case dispersion k and the effective reproduction number R_e. Increasing local transmission ($R_e > 1$) leads to greater uncertainty in the probability estimates. **Discussion:** When public health measures are effective, lower R_e (less transmission on average) and larger k (lower risk of superspreading) will be in effect and end-of-outbreak determinations can be declared with greater confidence. The application of end-of-outbreak probabilities can help distinguish between local extinction and low levels of transmission, and communicating these end-of-outbreak probabilities can help inform public health decision-making with relation to the appropriate use of resources.

Chapter 3: Correlation between times to SARS-CoV-2 symptom onset and secondary transmission undermines epidemic control efforts.

Methods: We used publicly available case data from Japan to reconstruct networks of transmission for different prefectures in Japan and identify infector-infectee pairs where exposure dates and symptom onset for the infector as well as dates of contact between the infector and

infectee were reported. From this data, we used doubly-interval censoring and copula methods to jointly estimate the generation interval and incubation period and obtain an estimate of correlation between the two epidemiological parameters. **Results:** We collected a dataset of 257 SARS-CoV-2 transmission pairs in Japan and jointly estimated the mean generation interval (3.7–5.1 days) and mean incubation period (4.4–5.7 days), taking into consideration sociodemographic and transmission characteristics. The generation interval and incubation period were positively correlated (Kendall's tau of 0.4–0.7). **Discussion:** Accounting for this dependence can improve the prediction of SARS-CoV-2 transmission in epidemic models as well as enhance evaluation of the effectiveness of public health interventions on disease control. However, as SARS-CoV-2 variants begin to dominate transmission in many countries, whether the correlation between the generation interval and incubation period would be weaker or stronger than has been presented here remains to be seen. Nonetheless, this study provides a basis for consideration of such correlation moving forward.

Conclusion: The COVID-19 pandemic has proven longer-lasting and more severe than most predicted in early 2020. Partially this is due to the large fraction of presymptomatic and asymptomatic transmission. Understanding of epidemiological parameters related to SARS-CoV-2 infection has been critical to the creation and fine-tuning of statistical models describing disease transmission, severity, and the impact of nonpharmaceutical interventions on control. The present dissertation has provided estimates for key epidemiological parameters of COVID-19 and provided guidance on a recommended quarantine period for COVID-19 cases at the beginning of the pandemic. In addition, it shows how epidemiological parameters are applied to end-of-outbreak estimation. Lastly, it elucidated the role of presymptomatic transmission in transmission and evaluated the effectiveness of symptom-based case isolation given the level of correlation between the incubation period and generation interval. There is much work yet to be done to determine epidemiological parameters related to vaccination and SARS-CoV-2 variants, but the work presented here will continue to inform COVID-19 models moving forward.

List of abbreviations

CDC	United States Centers for Disease Control and Prevention
CDF	Cumulative distribution function
CI/CrI	Confidence/credible interval
COVID-19	Coronavirus disease 2019
LOOIC	Leave-one-out information criterion
MCMC	Markov chain Monte Carlo
MERS	Middle East respiratory syndrome
MLE	Maximum likelihood estimation
PCR	Polymerase chain reaction
PDF	Probability distribution function
R_0	Basic reproduction number
R_e	Effective reproduction number
SARS	Severe acute respiratory syndrome
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
WAIC	Widely applicable information criterion
WHO	World Health Organization

Introduction

At the end of December 2019, information on a newly emerged disease began to appear on public health news feeds. While initially referred to as an atypical pneumonia of unknown etiology, by early January 2020 it was clear that this atypical pneumonia was a “Disease X”—a disease caused by an unknown pathogen with the potential to become epidemic—if not pandemic (Jung et al., 2020a; Mehand et al., 2018). Shortly thereafter, the pathogen of this Disease X was identified as a novel coronavirus and given the official name severe acute respiratory disease coronavirus 2 (SARS-CoV-2) while the disease it caused was termed coronavirus disease 2019 (COVID-19). Countries that had experienced outbreaks of other diseases caused by recently emerged human coronaviruses—severe acute respiratory syndrome (SARS), which emerged in late 2002, and Middle East respiratory syndrome (MERS), which emerged in 2012—were on high alert. However, much of the rest of the world simply kept a wary eye on the rapidly developing situation.

Early reports cast COVID-19 as primarily being the result of a single large zoonotic spillover event, posing limited risk to humans (Chinese National Health Commission (NHC), 2021). This was because many of the cases reported in December 2019 and early January 2020 were associated with a wet market in Wuhan, China—a place where animals (including some wild animals) and humans were in close proximity. However, evidence of that many infections had epidemiological consistency with continual human-to-human transmission rapidly accumulated (Nishiura et al., 2020a; Wu et al., 2020), and estimation of the epidemiological parameters governing the transmission dynamics of SARS-CoV-2 became a priority, in order to fully grasp the risk the novel virus posed to humans (Cowling and Leung, 2020).

Clues could be—and were—taken from coronavirus siblings SARS and MERS. The earliest models of SARS-CoV-2 transmission dynamics used the incubation period (time between infection and illness onset) from SARS and/or MERS to estimate the total number of cases and transmissibility of COVID-19 (Imai et al., 2020a, 2020b). However, any differences between the epidemiological parameters of COVID-19, SARS, and MERS could leading to under- or overestimation of epidemic speed/strength when used in transmission models (Gostic et al., 2020; Park et al., 2020b). As such models factored into policy decisions on timing of the

implementation and release of public health and social measures used to control disease spread (Ferguson et al., 2020), efforts were made to empirically estimate the epidemiological parameters for COVID-19. Such parameters included the incubation period, the serial interval—the time from symptom onset in an infected individual to symptom onset in a person they infect, as well as the generation interval—the time from infection of an infected individual to the time when they infect someone else.

Determining the incubation period of COVID-19 was a particularly urgent goal in the early phase of the pandemic—even before it was declared a pandemic—because the incubation period is used to inform a recommended length of quarantine (movement restriction) for people who may have been exposed to SARS-CoV-2. Quarantine to control the spread of the virus was urgently needed, as the epidemic had rapidly expanded in Wuhan in January 2020. Although travel to and from Wuhan was banned on January 23, and many international flights to and from mainland China were suspended or limited beginning in February (Chinazzi et al., 2020), these actions were too late to prevent domestic spread in countries that had already imported the virus. To address the need for evidence-informed quarantine policies, *Chapter 1* presents early estimates of the incubation period as well as other epidemiological parameters of COVID-19 and identifies a recommended quarantine period. By March 2020, most countries had restricted inbound travel and instituted quarantine policies for new arrivals as they grappled with controlling local transmission (Hale et al., 2021).

During the early part of 2020 it also became clear that COVID-19 was highly overdispersed, meaning that a small fraction of individuals were responsible for infecting the majority of new cases (Endo et al., 2020a). Japan, recognizing from experience with SARS that greater overdispersion results in a larger proportion of new cases being associated with case clusters (Knight et al., 2020), focused a portion of its COVID-19 control efforts specifically on the prevention and identification of clusters (Endo et al., 2020b; Oshitani, 2020). However, while such efforts had been successful with SARS (which disappeared a year and a half after it emerged), COVID-19 elimination proved difficult. In countries that did reach zero incident cases for an extended period of time (e.g., Taiwan, New Zealand), importation of cases later led to renewed domestic transmission (Akhmetzhanov et al., 2021; Baker et al., 2020). For countries with ongoing transmission, a great concern has been when to implement and lift public health

measures targeted towards disease control, as such measures consume resources and restrict economic activities. While many public health measures are implemented at the national or subnational level, locations and subpopulations associated with case clusters also face enhanced public health measures following an outbreak, and it is therefore beneficial to consider timing of lifting of restrictions related to COVID-19 clusters. In line with this consideration, *Chapter 2* leveraged previously estimated epidemiological parameters of COVID-19 to consider end-of-outbreak probabilities for individual case clusters.

As the pandemic continued it became clear that the large proportion of cases that were asymptomatic (Nishiura et al., 2020b) or infectious while presymptomatic (Tindale et al., 2020) is large part of why COVID-19 has proven so difficult to control. Asymptomatic and presymptomatic transmission indicate post-symptomatic case isolation efforts alone are insufficient to control transmission (Contreras et al., 2021; Lovell-Read et al., 2021). *Chapter 3* explored the association between the generation interval—the time between infection in an individual and when they infect another individual—and incubation period of COVID-19 transmission pairs detected in Japan to improve understanding of relationship between presymptomatic transmission and effectiveness of symptom-based case isolation.

Together, the three self-contained papers in this dissertation are unified around the theme of elucidating the epidemiological parameters of COVID-19 and considering how they can be applied to public health policies and models of transmission. As the world transitions to considering "endgame" scenarios for what has been a long pandemic response, keeping abreast of changes to how the virus and its host populations (predominantly humans) interact—and what this means in terms of the epidemiological parameters of COVID-19—will be vital.

Chapter 1: Early estimates of the incubation period and other epidemiological parameters of COVID-19

Background

As of 31 January 2020, mainland China reported 11791 confirmed cases of novel coronavirus (SARS-CoV-2) infections, causing 259 deaths. Initially, infections were thought to be due to zoonotic (animal-to-human) transmission, however recently published evidence (Li et al., 2020) and the exponential growth of case incidence show compelling evidence of human-to-human secondary transmission fueled by travel, with many cases detected in other parts of the world. This geographic expansion beyond the initial epicenter of Wuhan provides an opportunity to study the natural history SARS-CoV-2 infection, as these migration events limit the risk of infection to the time during which an individual traveled to an area where exposure could occur (Nishiura et al., 2007).

The incubation period is defined as the time from infection to illness onset. Knowledge of the incubation period of a directly transmitted infectious disease is critical to determine the time period required for monitoring and restricting the movement of healthy individuals (i.e. the quarantine period) (Lessler et al., 2009; Nishiura, 2009). The incubation period also aids in understanding the relative infectiousness of SARS-CoV-2 and can be used to estimate epidemic size (Nishiura et al., 2020c).

Time-delay distributions including dates of hospital admission (for treatment and/or isolation) and death also inform the temporal dynamics of epidemics. A published clinical study on the SARS-CoV-2 epidemic has already shown that the average time delay from illness onset to hospital admission is approximately 7 days (Huang et al., 2020), but this distribution has yet to be explicitly estimated. The time from hospital admission to death is also critical to avoidance of underestimation when calculating the case fatality risk (Donnelly et al., 2003). Using publicly available data of the ongoing epidemic of SARS-CoV-2 infected cases with known event dates, the present study aimed to estimate the incubation period and other time intervals that govern the interpretation of epidemiological dynamics of SARS-CoV-2 infections.

Methods

Epidemiological data

We retrieved information on cases with confirmed SARS-CoV-2 infection and diagnosis outside of the epicenter of Wuhan, China, based on official reports from governmental institutes. We aggregated the data directly from government websites or from news sites that quoted government statements. The data were collected in real time, and thus may have been updated as more details on cases became available.

Specifically, we collected the dates of exposure (entry and/or exit from Wuhan or dates of close contact with a Wuhan resident/known epidemic case), illness onset, earliest healthcare seeking related to infection, hospital admission (for treatment and/or isolation), and death. Cases included both residents from other locations who travelled to Wuhan, as well as individuals who lived, worked, or studied in Wuhan (hereafter: Wuhan residents) but were diagnosed outside of Wuhan and reported by the governments of the locations where SARS-CoV-2 infection was detected. We thus estimated the incubation period by (i) excluding Wuhan residents and (ii) including Wuhan residents. The former may be more precise in defining the interval of exposure, but the sample size is greater for the latter.

Statistical model

We used the dates of three critical points in the course of infection—symptom onset, hospital admission, and death—to calculate four time intervals: the times from (a) exposure to illness onset (i.e., incubation period), (b) illness onset to hospital admission, (c) illness onset to death, and (d) hospital admission to death. We used a doubly interval-censored (Reich et al., 2009) likelihood function to estimate the parameter values for these intervals, written as:

$$L(\Theta_g; \boldsymbol{D}) = \prod_i \int_{E_{L.i}}^{E_{R.i}} \int_{S_{L.i}}^{S_{R.i}} g(e) f(s-e) \mathrm{d}s \mathrm{d}e. \quad (1)$$

Here, in the case of (a) $g(.)$ is the probability density function (PDF) of exposure following a uniform distribution, and $f(.)$ is the PDF of the incubation period independent of $g(.)$. D represents a dataset among all observed cases i, and exposure and symptom onset fall within

lower and upper bounds (E_L,E_R) and (S_L,S_R). We fit the PDF $f(.)$ to lognormal, Weibull, and gamma distributions.

To address the selection bias in the dataset due to the continued growth of the outbreak (i.e., cases with shorter incubation periods are more likely to be included in the dataset), we also accounted for right truncation using the formula:

$$f'(s-e,e) = \frac{f(s-e)}{\int_0^{T-e} \frac{r\exp(-ru)}{1-\exp(-ru)} F(T-e-u)\mathrm{du}}. \quad (2)$$

Here, r is the exponential growth rate—estimated at 0.14 (Jung et al., 2020b)—T is the latest time of observation (31 January), and $F(.)$ is the cumulative density function of $f(.)$.

In both cases, we used Bayesian methods to infer parameter estimates and obtain credible intervals. We selected the best fit model by using widely applicable information criterion (WAIC). We also verified that the Bayesian estimates were in line with pointwise estimates derived by maximum likelihood estimation (MLE). As formulation of the likelihood with right truncation (1)–(2) contained the function f' and was dependent on both the time interval $(s-e)$ and time of exposure e, we generalized a previously obtained result for doubly interval-censored likelihood with$f'(s-e,e) \equiv f(s-e)$ (Donnelly et al., 2003).

The data were processed using R version 3.6.2, MLE was computed using Julia version 1.3, and the Markov chain Monte Carlo (MCMC) simulations were performed in Stan using cmdStan version 1.2.1 and StanHeaders version 2.21.0.1 (Stan Development Team, 2021).

Results

The ratio of male to female cases among living cases resembled that of Li et al. (2020), at 58%, with most 30–59 years of age (information missing for 9 cases). The deceased cases were more predominantly male (70%) and older (85% were 60 years of age or older). Table 1 shows estimates for the various time intervals without right truncation. For the incubation period estimates, the lognormal distribution provided the best fit to the data, both when excluding and including Wuhan residents. The mean incubation period was estimated at 4.6 days (95% credible

interval [CI]: 3.8, 5.7) when excluding Wuhan residents (n=52) and 5.0 days (95% CI: 4.4, 5.6) when including Wuhan residents (n=158).

The median time from illness onset to hospital admission was estimated at 3.8 days (95% CI: 2.9, 5.1) among living cases and 6.2 days (95% CI: 5.0, 7.8) among deceased cases using the gamma distribution, which provided the best fit for both sets of data. Figure 1A shows the corresponding PDFs. Time from symptom onset and hospital admission to death best fit lognormal and Weibull distributions, respectively, presented in Figures 2B and 2C. The mean time from onset to death was 14.6 days (95% CI: 12.5, 17.0) and from hospital admission to death was 8.6 days (95% CI: 6.8, 10.8).

Table 2 shows estimates for the best-fit model for each interval when accounting for right truncation. The mean incubation period was 5.4 days (95% CI: 4.1, 7.4) when excluding Wuhan residents—slightly larger than the estimate without right truncation. The mean estimate for illness onset to hospital admission was 9.7 days (95% CI: 5.4, 17.0) for living cases and 6.6 days (95% CI: 5.2, 8.8) for deceased cases, with the former nearly 2.5 times the length of its untruncated version. Onset to death and hospital admission to death were likewise longer than their non-truncated counterparts, at 19.9 days (95% CI: 14.9, 29.0) and 12.8 days (95% CI: 8.6, 20.2), respectively.

Figure 2 shows the cumulative distribution function of the incubation period with and without right truncation. The 5th and 95th percentiles are shown in addition to the median. The 95th percentiles were estimated at 11.5 days (95% CI: 8.8, 16.3) and 10.4 days (95% CI: 8.9, 12.5) for non-truncated data excluding and including Wuhan residents and 14.4 days (95% CI: 10.1, 24.3) when applying right truncation and excluding Wuhan residents. The respective median values for these CDFs were 3.5 days (95% CI: 2.7, 4.4), 4.3 days (95% CI: 3.8, 4.9), and 3.7 days (95% CI: 2.8, 4.8).

Table 1. Incubation period and other time-delay distributions for COVID-19 outbreak cases reported in January 2020.

	Incubation period excluding WR (days)	Incubation period including WR (days)	Onset to hospitalization, living (days)	Onset to hospitalization, deceased (days)	Onset to death (days)	Hospitalization to death (days)
Number of cases	52 cases	158 cases	155 cases	34 cases	34 cases	39 cases
Lognormal						
Mean	4.6 (3.8, 5.7)	5.0 (4.4, 5.6)	3.8 (2.9, 5.1)	6.2 (5.0, 7.8)	14.6 (12.5, 17.0)	8.6 (6.8, 10.8)
SD	3.9 (2.5, 6.4)	2.9 (2.2, 3.8)	8.3 (4.9, 14.2)	4.3 (2.9, 6.6)	6.7 (4.9, 9.4)	6.7 (4.6, 10.3)
5%	1.1 (0.6, 1.7)	1.8 (1.4, 2.3)	0.2 (0.1, 0.3)	1.9 (1.3, 2.5)	6.5 (4.9, 7.9)	2.2 (1.4, 3.0)
Median	3.5 (2.7, 4.4)	4.3 (3.8, 4.9)	1.6 (1.2, 2.0)	5.1 (4.1, 6.3)	13.2 (11.3, 15.2)	6.7 (5.3, 8.3)
95%	11.5 (8.8, 16.3)	10.4 (8.9, 12.5)	13.6 (10.1, 19.2)	13.9 (10.7, 19.6)	26.8 (22.3, 34.3)	20.5 (15.7, 28.8)
99%	18.8 (13.1, 31.2)	14.9 (12.1, 19.2)	33.0 (22.4, 52.2)	21.0 (15.1, 32.6)	36.0 (28.7, 49.1)	32.6 (23.4, 50.4)
WAIC	**360.0**	944.7	693.8	183.9	**221.9**	240.1
Weight	**0.55**	1.00	0.00	0.46	**0.84**	0.14
Weibull						
Mean	4.7 (3.7, 5.8)	5.1 (4.5, 5.8)	3.3 (2.7, 4.1)	6.2 (5.0, 7.7)	14.8 (12.3, 17.5)	8.6 (7.0, 10.4)
SD	4.2 (3.1, 6.1)	3.2 (2.7, 3.8)	4.2 (3.2, 5.7)	4.0 (3.1, 5.3)	7.6 (6.2, 9.8)	5.4 (4.2, 7.2)
5%	0.4 (0.1, 0.8)	1.0 (0.6, 1.4)	0.1 (0.0, 0.1)	1.1 (0.6, 1.9)	3.9 (2.3, 5.7)	1.6 (0.8, 2.5)
Median	3.6 (2.5, 4.6)	4.6 (3.9, 5.3)	1.8 (1.4, 2.3)	5.5 (4.2, 7.0)	13.9 (11.3, 16.6)	7.7 (5.9, 9.5)
95%	12.7 (10.0, 17.6)	11.0 (9.7, 12.8)	11.4 (9.1, 14.9)	13.5 (11.1, 17.8)	28.4 (24.3, 34.7)	18.5 (15.4, 23.8)
99%	18.4 (13.9, 28.5)	14.3 (12.3, 17.1)	19.5 (14.9, 27.4)	17.5 (14.1, 24.3)	35.0 (29.5, 44.3)	24.1 (19.5, 32.6)
WAIC	373.1	984.9	663.1	185.4	231.2	**236.1**
Weight	0.001	0.00	0.00	0.21	0.02	**0.45**
Gamma						
Mean	5.0 (4.1, 5.9)	5.1 (4.5, 5.7)	3.2 (2.6, 4.0)	6.3 (5.0, 7.9)	14.9 (12.3, 18.1)	8.7 (7.0, 10.8)
SD	3.5 (2.7, 4.6)	3.1 (2.6, 3.8)	4.2 (3.3, 5.4)	4.3 (3.2, 5.9)	8.8 (6.8, 11.6)	6.0 (4.6, 8.1)
5%	0.9 (0.5, 1.4)	1.2 (0.8, 1.6)	0.0 (0.0, 0.1)	1.2 (0.6, 1.9)	4.0 (2.5, 5.5)	1.6 (0.9, 2.5)
Median	4.2 (3.3, 5.0)	4.4 (3.8, 5.0)	1.7 (1.2, 2.2)	5.3 (4.1, 6.7)	13.2 (10.7, 16.0)	7.2 (5.7, 9.1)
95%	11.6 (9.5, 14.6)	11.0 (9.6, 12.9)	11.4 (9.3, 14.7)	14.3 (11.4, 19.3)	31.3 (25.7, 39.8)	20.1 (16.1, 26.3)
99%	16.1 (13.0, 21.1)	14.9 (12.8, 18.0)	19.1 (15.1, 25.1)	19.8 (15.4, 27.5)	41.8 (33.7, 54.8)	27.9 (21.9, 37.7)
WAIC	361.1	974.9	**528.4**	**183.3**	227.8	236.1
Weight	0.45	0.00	**1.00**	**0.33**	0.14	0.42

WR: Wuhan residents. SD: standard deviation. WAIC: widely applicable information criterion. Hospitalization (for living cases) includes isolation.

Table 2. Right-truncated incubation period and other time-delay distributions for COVID-19 outbreak cases reported in January 2020 applied to the lognormal distribution.

Lognormal	Incubation period excluding WR (days)	Onset to hospitalization, living (days)	Onset to hospitalization, deceased (days)	Onset to death (days)	Hospitalization to death (days)
Mean	5.4 (4.1, 7.4)	9.7 (5.4, 17.0)	6.6 (5.2, 8.8)	19.9 (14.9, 29.0)	12.8 (8.6, 20.2)
SD	5.7 (3.0, 11.6)	35.2 (12.9, 84.5)	4.8 (3.0, 8.5)	11.4 (6.5, 21.6)	12.3 (6.3, 24.5)
5%	1.0 (0.5, 1.6)	0.2 (0.1, 0.3)	1.8 (1.2, 2.5)	7.3 (5.6, 9.3)	2.5 (1.7, 3.5)
Median	3.7 (2.8, 4.8)	2.6 (1.9, 3.8)	5.3 (4.2, 6.7)	16.8 (13.4, 23.7)	9.0 (6.7, 13.1)
95%	14.4 (10.1, 24.3)	35.1 (20.5, 65.0)	14.9 (11.1, 23.8)	39.1 (27.7, 69.3)	32.4 (20.6, 59.3)
99%	25.3 (15.6, 52.5)	102.8 (52.8, 218.4)	23.0 (16.0, 42.0)	55.3 (36.6, 109.0)	55.0 (31.9, 114.7)

WR: Wuhan residents. SD: standard deviation. Hospitalization (for living cases) includes isolation.

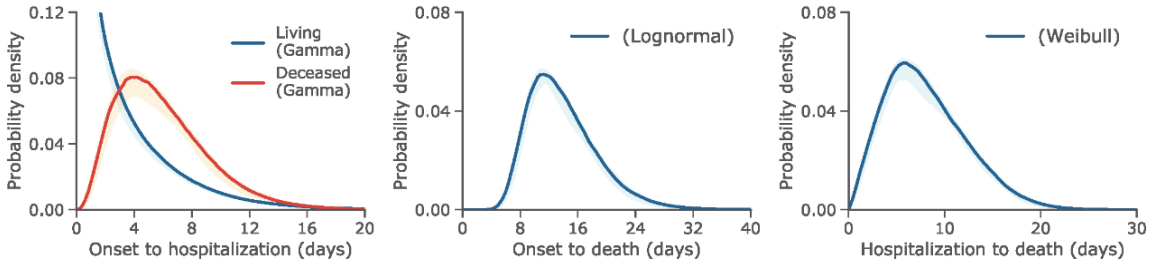

Figure 1. Probability distributions of time from onset or hospital admission to hospital admission or death for COVID-19 cases reported through 31 January 2020.

(A) Probability density of the time from illness onset to hospitalization in days set to the best-fit gamma distribution. (B) Probability density of the time from illness onset to death in days set to the best-fit lognormal distribution. (C) Probability density of the time from hospital admission to death in days set to the best-fit Weibull distribution.

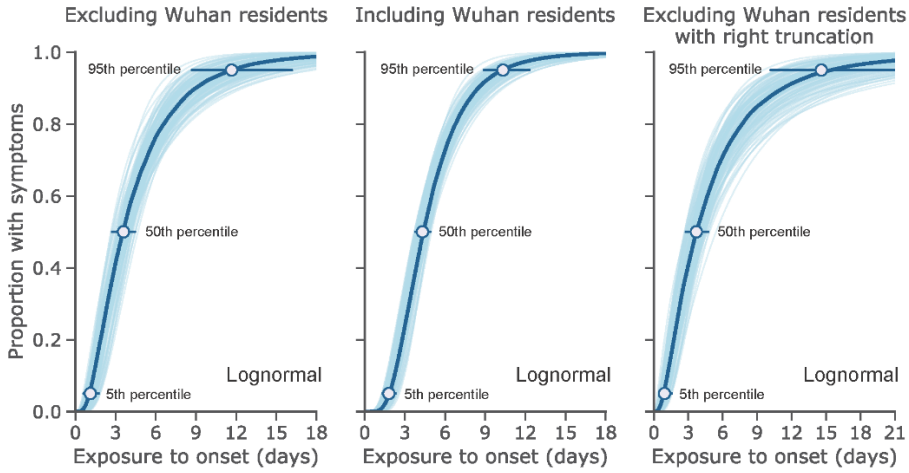

Figure 2. Estimated cumulative distribution for the incubation period of SARS-CoV-2 infections from outbreak cases reported through 31 January 2020.

Data are from public case reports. Left and center: non-truncated estimates excluding (n=52) and including (n=158) Wuhan residents. Right: right-truncated estimates excluding Wuhan residents (n=52).

Discussion

The present study advances the public discussion on SARS-CoV-2 infections by presenting explicit estimations of the incubation period and other epidemiologic characteristics using publicly available data. Our estimated mean incubation period of approximately 5 days is comparable to known mean values of the incubation period for severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) (Donnelly et al., 2003; Lessler et al., 2015; Virlogeux et al., 2016), as well as other recent estimates of the incubation period for COVID-19 (Backer et al., 2020). In addition to empirically showing the comparability of SARS-CoV-2 to other disease-causing coronaviruses, the present study has also shown that the 95th percentile of the incubation period is around 10–14 days, indicating that a 14-day quarantine period would largely ensure the absence of disease among healthy exposed individuals.

Wuhan residents have a less precisely defined exposure period compared to travelers and secondary cases from known human to human transmission events. However, our calculations have shown that adding more cases to the dataset even with uncertainty reduces both the variance of the estimates and selection bias, improving the fit of the mean of the distribution. Our estimates are in agreement with the report of Li et al. (2020). A recent study by Backer et al. (Backer et al., 2020) noted a similar finding in their analysis of the incubation period for 88 cases (including 63 Wuhan residents). However, the estimates of Backer et al. for the model that included Wuhan residents were subject to overestimation as the lower bounds for Wuhan residents—who had unknown left exposure dates—were fixed in their analysis. In contrast, we considered the left exposure dates for Wuhan residents as parameters to be fitted. Notably, our results demonstrated the overall benefit of using additional case data even when some of their exposure values were unknown.

The time from illness onset to death is also comparable to SARS (Lessler et al., 2015), and the lengthy 15–20-day mean delay indicates that crude estimation of the ratio of the cumulative number of deaths to that of confirmed cases will tend to result in underestimation of the case fatality risk, especially during the early stage of epidemic spread. During the SARS epidemic in Hong Kong, 2003, the time from illness onset to hospitalization was shown to have shortened as a function of calendar time, gradually reflecting the effects of contact tracing (Donnelly et al.,

2003). It remains to be seen if this will be the case for COVID-19 as well. The time delay distribution between illness onset and hospital admission may also be negatively associated with the basic reproduction number, i.e., the average number of secondary cases generated by a single primary case in a fully susceptible population.

The median time from illness onset to hospital admission was approximately 4 days among cases not known to be deceased at time of case report, and 6 days among cases reported as deceased. The reasons for this difference are not altogether clear. However, the living cases include persons who are isolated—in some cases more for reducing transmission than for treatment purposes—while all deceased cases were hospitalized. In addition, deceased cases for whom information was available had onset dates closer to the beginning of the outbreak compared to the living cases, who mostly had onset in the latter two-thirds of January. The time delay distributions from onset to hospital admission for cases reported later in the epidemic, when there was more widespread recognition of the virus and a more prevalent social imperative for those with symptoms to seek healthcare, may differ from those of early cases.

Several limitations of the present study exist. First, the dataset relies on publicly available information that is not uniformly distributed (i.e. collected from various sources), and therefore the availability of dates relevant to our analyses are limited to a small, selective sample that is not necessarily generalizable to all confirmed cases. Moreover, given the novelty of the COVID-19, it is possible that illness onset and other event data were handled differently between jurisdictions (e.g., was onset the date of fever or date of dyspnea?). Second, our data include very coarsely date intervals with some proxy dates used to determine the left and/or right hand dates of some intervals. Third, as the sample size was limited the variance is likely to be biased. Fourth, we were not able to examine heterogeneity of estimates by different attributes of cases, e.g., severity of disease (Virlogeux et al., 2016). Lastly, as we only have information on confirmed cases, there is a bias towards more severe disease—particularly for earlier cases.

This study presents estimates of epidemiological characteristics of SARS-CoV-2 infections that are key parameters studies on incidence, case fatality, and epidemic final size, among other possibilities (Jung et al., 2020b; Nishiura et al., 2020c). From the 95th percentile estimate of the incubation period we found that the length of quarantine should be at least 14 days, and stress

that the 15–19-day time delay from illness onset to death must be addressed when estimating COVID-19 case fatality risk. This study was made possible only through open sharing of case data from China and other countries where cases were diagnosed. Continued communication of dates and other details related to exposure and infection is crucial to furthering scientific understanding of the virus, the infections it causes, and preventive measures that can be used to contain and mitigate epidemic spread.

Supplementary materials

Estimation of the time interval distribution using doubly interval-censored likelihood

Here we consider the fit of the incubation period as the time interval between the windows of exposure *"e"* and symptom onset *"s"* (written in lower or upper case). This approach can be applied to other time intervals, such as the interval between illness onset and death, illness onset and hospitalization, etc. The doubly interval-censored likelihood holds the following form:

$$L(\boldsymbol{\Theta} \mid \boldsymbol{D}) = \prod_i \int_{E_{L,i}}^{E_{R,i}} \int_{S_{L,i}}^{S_{R,i}} g(e) f'(s-e, e)\, \mathrm{d}s\, \mathrm{d}e, \qquad \text{(S1)}$$

where *L* refers to the left-hand (lower) value of a window of exposure or symptom onset, and *R* is the right-hand (upper) value; *i* is the index over all available data record of pairs $(S_{L,i}, S_{R,i})$ and $(E_{L,i}, E_{R,i})$ with the assumption that at least $S_{R_i} > E_{L_i}$. A previously published work by Reich and colleagues (2009) considered the same problem but under the condition of a uniformly distributed exposure distribution and $f'(s-e, e) \equiv f(s-e)$. The results of Reich and colleagues can be altered to include a more general form of the function under the integral. Here, we describe how this can be done.

The main difficulty in calculating the likelihood (S1) is addressing the double integral over the rectangular area $[S_{L,i}, S_{R,i}] \times [E_{L,i}, E_{R,i}]$:

$$\mathcal{I}_i = \int_{E_{L,i}}^{E_{R,i}} \int_{S_{L,i}}^{S_{R,i}} g(e) f'(s-e, e)\, \mathrm{d}s\, \mathrm{d}e. \qquad \text{(S2)}$$

We notice that the integral $\mathcal{I}_i$ can be transformed to the repeated integral after changing the variable $s' = s - e$, and subsequently calculated using Figure 1.

Figure 1A shows the situation when $S_{L,i} > E_{R_i}$. In this case:

$$\begin{aligned}\mathcal{I}_i &= \int_{E_{L,i}}^{E_{R,i}} \mathrm{d}e\; g(e) \int_{S_{L,i}-e}^{S_{R,i}-e} f'(s', e)\, \mathrm{d}s' \\ &= \int_{E_{L,i}}^{E_{R,i}} g(e) \left(F'\left(S_{R,i} - e, e\right) - F'\left(S_{L,i} - e, e\right) \right) \mathrm{d}e,\end{aligned} \tag{S3}$$

where F' is a cumulative distribution function (CDF) of f'.

Figure 1B shows the situation when $E_{R_i} > S_{L,i} > E_{L,i}$. In this case:

$$\begin{aligned}\mathcal{I}_i = &\int_{E_{L,i}}^{S_{L,i}} g(e) \left(F'\left(S_{R,i} - e, e\right) - F'\left(S_{L,i} - e, e\right) \right) \mathrm{d}e \\ &+ \int_{S_{L,i}}^{E_{R,i}} g(e) F'\left(S_{R,i} - e, e\right) \mathrm{d}e.\end{aligned} \tag{S4}$$

Lastly, Figure 1C shows the situation when $E_{L_i} > S_{L,i}$. In this instance:

$$\mathcal{I}_i = \int_{E_{L,i}}^{E_{R,i}} g(e) F'\left(S_{R,i} - e, e\right) \mathrm{d}e. \tag{S5}$$

We note that the doubly interval-censored likelihood with truncation admits the function f' to be separable on its arguments: $f'(s - e, e) = f(s - e)h(e)$. This means that the CDF: $F'(s - e, e) = F(s - e)h(e)$ contains the CDF of the function $f(s - e)$ and can be analytically derived for all three gamma, lognormal, and Weibull distributions considered in the main text.

Our numerical simulations confirmed that calculation of the likelihood using the formulas (S3)–(S5) lead to results that are identical to those produced using the R package coarseDataTools developed by Reich and colleagues (Reich, 2019).

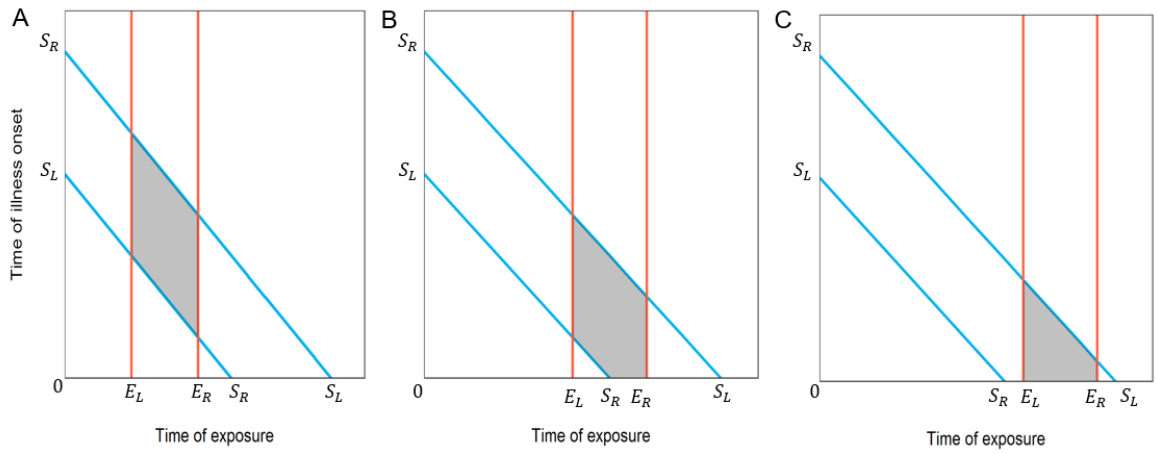

Figure S1. Illustration of the time interval distribution using doubly interval-censored likelihood.

The figure depicts how doubly interval-censored likelihood was calculated when A) $\boldsymbol{S_{L,i} > E_{R_i}}$, B) $\boldsymbol{E_{R_i} > S_{L,i} > E_{L,i}}$ and C) $\boldsymbol{E_{L_i} > S_{L,i}}$. The blue lines and red lines indicate the lower or upper bounds of illness onset and exposure, respectively. The gray shaded area represents the integral in each situation.

Estimation of the time interval distributions using a Bayesian framework.

Here, we estimate parameters of time interval distributions for non-truncated likelihood using Markov chain Monte Carlo (MCMC) simulations. We used Stan packages with implementation of No U-Turn Sampler (NUTS). Specifically, each distribution is specified by the following prior distributions:

- The lognormal distribution is defined by the logarithms of the mean and the logarithm of the standard deviation. Both follow the standard normal distribution (with mean set at zero, and standard deviation equal to one).

- The Weibull distribution is defined by the logarithm of the mean and the logarithm of the shape parameter. Both follow the standard normal distribution.

- The gamma distribution is defined by the shape and inverse scale parameter. The former follows the normal distribution with the mean three and standard deviation five and is constrained to positive values. The latter follows a Cauchy distribution with location parameter set to zero and scale parameter equal to 5.0.

We followed general recommendations of the Stan developer community for the choice of prior distributions (Stan Development Team).

Additionally, we specified the priors for the times of exposure and illness onset (if necessary, relative to the other time interval distributions) for each case i:

$$e_i = E_{L,i} + \left(E_{R,i} - E_{L,i}\right)\tilde{e}_i, \qquad s_i = \hat{S}_{L,i} + \left(S_{R,i} - \hat{S}_{L,i}\right)\tilde{s}_i,$$

where: $\hat{S}_{L,i} = e_i$ if $e_i > S_{L,i}$, and $\hat{S}_{L,i} = S_{L,i}$, otherwise, and $\tilde{e}_i$ and $\tilde{s}_i$ are a pair of standardized variables:

$$\tilde{e}_i \sim \text{normal}(\text{mean} = 0.5, \text{SD} = 0.5), \qquad \tilde{s}_i \sim \text{normal}(\text{mean} = 0.5, \text{SD} = 0.5),$$

on the interval between zero and one.

The non-truncated loglikelihood written in terms of the Stan language was defined as:

$$\log L(\boldsymbol{\Theta}) = \sum_i \text{distibution\_lpdf}(s_i - e_i|\boldsymbol{\Theta}),$$

where distibution_lpdf is either the function lognormal_lpdf, weibull_lpdf, or gamma_lpdf; "lpdf" is shorthand for the logarithm of the probability distribution function.

We also used a slight variation of this model for the incubation period as to include censored likelihood to account for unknown lower bound of the exposure period.

In each case, we ran four chains of MCMC simulations consisting of 5,000 iterations for the simulation phase and 10,000 used for the tuning phase. The convergence of the obtained traces was verified visually, and we also explored the r-hat values and effective sample size. For example, Figure 2 shows the traces obtained for the fit of the incubation period to a lognormal distribution, which was the best-fit model as seen in Table 1 of the main text.

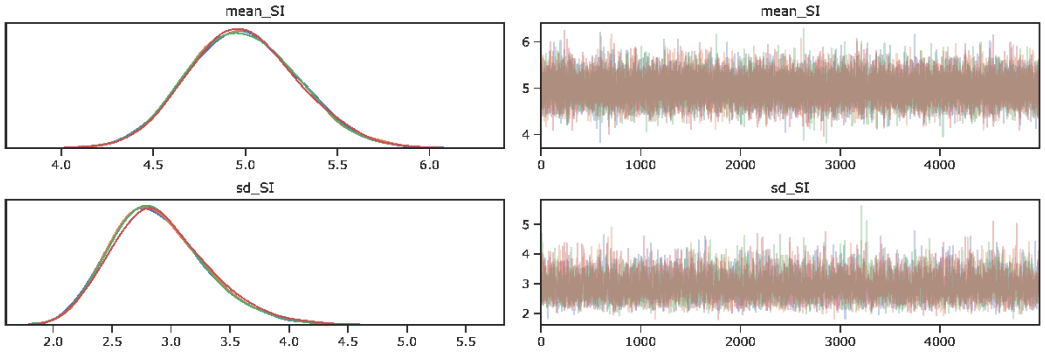

Figure S2. The density plots and traces for the mean and standard deviation of the lognormal distribution when fitted to the dataset of incubation periods including Wuhan residents.

Data cleaning rules implemented for the various time intervals

Incubation period:

For incubation periods of non-Wuhan residents (exposure type !="Lives/works/studies in Wuhan"):

- Only include cases with travel to Wuhan, contact with Wuhan residents, or contact with a case.
- Must have a left exposure date (EL).
- One of onset, date first sought healthcare, or date hospitalized/isolated must be larger than left exposure date.
- For calculating ER and SL, must have right exposure date or onset date.
- For calculating SR, must have onset date or date first sought healthcare or date hospitalized/isolated or date of report.
- If left exposure (EL) is the same as the right exposure (ER)—which was the case for some people with day trips to Wuhan or contact with cases—we set ER to EL + 1 to provide a "day" of exposure.
- If right exposure date was not available or was greater than date of onset, ER was set to date of onset.
- Left onset (SL) is the date of onset (if available), otherwise set to the day after ER.
- Right onset (SR) is the day after date of onset if available, otherwise set to the date first sought healthcare, date hospitalized/isolated, or date reported as case, in that order.

For incubation period including Wuhan residents, the only difference in methods is that we do not require left exposure date for Wuhan residents. If left exposure date is NA, they must have exposure type "Lives-works-studies in Wuhan" or are excluded. Then, if the left exposure date is missing and they are a Wuhan resident, E_L is set to 1 December 2019, which was the first date of onset for a reported case from Wuhan (Li et al., 2020).

Onset to hospitalization:

- Calculated separately for deceased cases (almost all Hubei residents) and cases diagnosed outside of Hubei Province:
- Must have date of onset.
- Must have date of hospitalization or date of report (death for death dataset).
- Left onset (EL) is the date of onset.
- Right onset (ER) is the date of onset + 1.
- Left hospitalization/isolation (SL) is date of hospitalization, if available, otherwise date first sought healthcare, otherwise date of onset + 1.
- Right hospitalization/isolation (SR) is date of hospitalization + 1 if available, otherwise date of report (for export cases) or death (for deceased cases).

Onset to death:

- Must have date of onset.
- Must have date of death.
- Left onset (EL) is the date of onset.
- Right onset (ER) is the date of onset + 1.
- If time of death is known, left date of death (SL) is the date of death + time of death - 1 hour, otherwise date of death - 1 day.
- If time of death is known, right date of death (SR) is the date of death + time of death, otherwise date of death.

Hospitalization to death:

- Must have date of hospitalization or onset.
- Must have date of death.
- Left hospitalization/isolation (SL) is date of hospitalization, if available, otherwise date first sought healthcare, otherwise date of onset + 1.

- Right hospitalization/isolation (SR) is date of hospitalization + 1 if available, otherwise date of death - 1.
- If time of death is known, left date of death (SL) is the date of death + time of death - 1 hour, otherwise date of death - 1 day.
- If time of death is known, right date of death (SR) is the date of death + time of death, otherwise date of death.

Chapter 1 was originally published as:

Linton NM, Kobayashi T, Yang Y, et al. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *J Clin Med*. 2020;9(2)538. Minor formatting modifications and edits have been made for the dissertation.

Chapter 2: Localized end-of-outbreak determination for coronavirus disease 2019 (COVID-19): examples from clusters in Japan

Background

Control of COVID-19 is difficult, and timely, localized information regarding which outbreaks are growing and which are likely to end is needed to inform control measures. Here, we explain a method to estimate end-of-outbreak probabilities at localized levels, allowing for evidence-based decision-making around the scaling-back of public health and social response measures in real-time. We demonstrate the applicability of this method in relation to clusters of cases using several examples from Japan.

Early research into the transmission dynamics of COVID-19 indicated that spread of the causal virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was highly overdispersed (Endo et al., 2020a; Lloyd-Smith et al., 2005). The degree of overdispersion is quantified by dispersion parameter k, which describes the variance in the distribution of the number of secondary cases infected by a typical primary case. Lower values of k represent greater variance and thereby a propensity towards superspreading—the phenomenon by which some infected individuals transmit a pathogen to large numbers of secondary cases, while most do not infect others (Lloyd-Smith et al., 2005). Larger values of k indicate less dispersion, and a Poisson distribution is obtained as a special case of the negative binomial distribution as $k \to \infty$ (Lloyd-Smith, 2007). During the pandemic, k for COVID-19 has been reported in the range of 0.1–0.6 (Bi et al., 2020; Endo et al., 2020a; Riou and Althaus, 2020; Tariq et al., 2020), indicating that most secondary infections are caused by a small fraction of primary cases, and therefore that superspreading events can fuel disease transmission.

In Japan, the propensity of COVID-19 towards superspreading was addressed by including the prevention, detection, and suppression of clusters—groups of cases linked by a common place and time—as a key component of national response (Oshitani and The Expert Members of the National COVID-19 Cluster Taskforce at the Ministry of Health Labour and Welfare of Japan, 2020). Additional public health and social measures took into consideration commonalities in transmission settings between clusters (Furuse et al., 2020)—most prominently in the form of a

nationwide messaging campaign encouraging residents to avoid the "Three Cs": closed spaces with poor ventilation, crowded places, and close-contact situations (Oshitani, 2020). Although Japan had relatively low levels of epidemic growth and fatalities reported during the first six months of the pandemic compared to other developed nations in Europe and North America, it faced a resurgence of cases during the summer months.

Current World Health Organization (WHO) guidance focuses on assigning levels of risk based on epidemiological, health system, and surveillance criteria from trends and other descriptive data to inform the loosening and re-tightening of response activities (World Health Organization (WHO), 2020a). The guidance indicates that an outbreak can be considered controlled if the effective reproduction number R_e—the average number of secondary cases produced per primary case in the presence of interventions and immune individuals in a given time period—is maintained below the threshold value of 1 for at least two weeks. While this distinguishes well between epidemic growth and decline, this method of surveillance does not differentiate between whether an outbreak will end entirely or will continue as stuttering chains of transmission (i.e. $0 < R_e < 1$) before potentially resurging ($R_e > 1$) (Blumberg and Lloyd-smith, 2013). Other COVID-19 guidance that address end-of-outbreak declarations relate to the incubation period-based guidelines used for other directly transmittable diseases such as measles and Ebola virus disease. For those diseases, two times the maximum incubation period—the time from infection to symptom onset—since the last possible date of exposure to a source of infection within the outbreak is used to determine the time until the end of an outbreak can be declared (Public Health Agency of Canada, 2020; World Health Organization (WHO), 2020b). Despite widespread use, this method has previously been proven flawed, and is particularly vulnerable when local surveillance systems are weak (Lee and Nishiura, 2017; Thompson et al., 2019).

An alternative statistically-based method is to leverage parameters that describe transmission dynamics underlying the epidemic curve to estimate the probability that one or more cases will be reported going forward in time (Nishiura et al., 2016). This more rigorous basis for end-of-outbreak determination can also be adjusted to account for additional factors such as case underascertainment and changing transmission dynamics. Including this statistical analysis adds to the information available to politicians and public health officials when making decisions

regarding response needs. Here, we present examples where this method was used to assess end-of-outbreak probabilities for clusters in Japan during the first wave of the pandemic.

Methods

Data collection and cluster selection

Epidemiological data were collected from the official case reports published online by reporting jurisdictions (prefectures and some cities) within Japan. In some instances, the data were supplemented and additionally verified using information from press briefings from the reporting jurisdictions. The collected data include date of onset, date of report, and linkage information used to inform grouping of cases into clusters. A dataset including dates and cluster information are included in the supplementary materials.

For this study, four clusters representing different social contexts and time points during the epidemic were selected. The first cluster occurred early in the pandemic in Aichi Prefecture with the first case reported on 14 February 2020 (Ministry of Health Labour and Welfare (MHLW), 2020). It describes several chains of transmission related to fitness gym use and social contact between cases. The second cluster was identified in Kyoto Prefecture in March and was initiated by importation of cases exposed to SARS-CoV-2 in Europe. Subsequent domestic transmission occurred during a series of events attended by university students and other contacts. The third and fourth clusters were linked to nosocomial transmission in a medical facility and a senior care facility in Hokkaido Prefecture with cases reported during April and May, respectively. Clusters include all secondary cases arising from chains of transmission linking back to the original sources of common exposure.

Statistical model

Each cluster is given by the epidemic curve represented at the time of report *t* with onset dates $t_i \leq t$ for an epidemiologically defined group of cases $i = \{1, \dots, M\}$. The probability that one or more new cases $X(t)$ will be reported on day t is written as follows (Nishiura et al., 2016):

$$\text{Prob}(X(t) > 0) = 1 - \prod_{i=1}^{M} \sum_{y=0}^{\infty} p_y [F(t - t_i)]^y, \quad \textbf{(3)}$$

where p_y is the probability that *y* secondary cases arise from a given primary case *i* following a negative binomial distribution with the mean R_e and variance $R_e(1 + \frac{R_e}{k})$, with *k* the dispersion parameter.

The function $F(.)$ defines the cumulative distribution function (CDF) of the serial interval *f* which is backprojected using the time delay from illness onset to report h, and defined by the convolution:

$$F(t - t_i) = \sum_{s=2}^{t-t_i} \sum_{\tau=1}^{s-1} f(s - \tau) h(\tau). \quad (4)$$

Outbreak extinction is determined once the estimate for the probability of observing one or more additional cases $\text{Prob}(X(t) > 0)$ comes below a given threshold. The day the probability estimate drops below the threshold is in effect the day the outbreak would be declared over. Selection of a threshold value depends on whether the goal is to minimize the observation period (higher threshold) or minimize risk that undetected cases may exist and become detected following the end of outbreak declaration (lower threshold). Here, we examine a 5% threshold, which translates to a 5% risk that the end-of-outbreak declaration would be preemptive and case(s) would be detected following the declaration (Lee and Nishiura, 2019).

Missing dates of illness onset

For some cases, date of onset was unavailable. Either the case did not give permission for disclosure, the information was not collected, or the case was asymptomatic at the time of detection and did not later report symptoms. The latter scenario may therefore represent either pre-symptomatic cases with no follow-up report or cases who were completely asymptomatic until recovery.

We approached missing dates of illness onset in two ways. First, we excluded cases with no reported onset date and calculated the probability of new cases solely based on the existing epidemic curve of illness onsets, hereafter referred to as the "reported" dataset. Second, we sampled the reporting delays and subtracted their values from the known report dates for all cases with no available date of illness onset to obtain a proxy onset date, hereafter referred to as the "imputed" dataset.

Underascertainment of cases

Despite efforts to obtain high quality surveillance data through contact tracing and testing, it is still likely that cases remain underascertained. Previous reports have suggested 9.2–44.4% case ascertainment using data on Japanese evacuees from the original epicenter of Wuhan, China and laboratory testing conducted in Japan during January and February 2020, respectively.(Nishiura et al., 2020d; Omori et al., 2020) For clusters with relatively stable and captive populations (e.g., medical centers and senior homes) ascertainment is likely to be higher due to intensive contact tracing on a focused population, though as chains of transmission move away from the common exposure setting, ascertainment will approach that of the general population. Clusters based on social contact linked to >1 common exposure setting (e.g., multiple restaurants, gyms, or other venues) are more likely to have lower levels of case ascertainment, though in Japan it is expected that ascertainment for cases related to a cluster would be higher than for the general population due to targeted case finding.

To address likely case underascertainment, we sampled from a binomial distribution with probability of success $p = 1 - q$, where q is the ascertainment rate. We assume that underascertained cases could only exist within one serial interval—i.e. 5 days (Nishiura et al., 2020e)—from the date of onset of the last reported case because the contact tracing team is unlikely to miss cases from two consecutive generations. A number of unreported cases u_τ at day τ can be then inferred using the following observational model:

$$\begin{aligned} & i_\tau \sim \text{Binomial}(\text{size} = u_\tau + i_\tau, \text{prob} = p), \\ & 0 \leq u_\tau \leq U, \qquad \tau \leq \max(t_i) + 5, \end{aligned} \tag{5}$$

where U is a maximally possible number of unreported cases that was assigned to 50 in our simulations.

Parameter selection and statistical analysis

The model was applied using data on the epidemic curves for individual clusters, accounting for missing dates of onset and varying levels of case ascertainment as described previously. The applied parameter values are shown in Table 3. R_e was explicitly varied between 0.5, 1.5, and 3, although an estimate of the local time-varying effective reproduction number (R_t) would be a sensible option when conducting analyses in real time. Otherwise, we accounted for parameter uncertainties via resampling. Estimates for k were drawn from a positive half-normal distribution using mean and SD from published studies (Bi et al., 2020; Tariq et al., 2020; Zhang et al., 2020b). Other distributions were also resampled from their empirical distributions. The serial interval distribution was previously reported by Nishiura et al. (Nishiura et al., 2020e) and the reporting delay was estimated from all for cases reported in Japan through the end of May using doubly interval-censored methods described elsewhere (Linton et al., 2020; Nishiura et al., 2020e). The reporting delay was estimated from cases reported through the end of May to coincide with the time interval during which cases for the four clusters were reported. We implemented the analysis using R 4.0.3 and CmdStan 2.26.1 (R Development Core Team, 2019).

Ethical considerations

The present study analyzed publicly available data which were already de-identified upon press release. The present study was approved by Medical Ethics Board of the Graduate School of Medicine, Kyoto University (R2673).

Results

Characteristics of the four clusters are shown in Table 4. The Aichi fitness gyms cluster had the smallest number of cases (n=40) while the Hokkaido senior care facility cluster was the largest (n=94). The average age was lowest for the Kyoto cluster, which was associated with parties attended primarily by university students, while the average age was highest for the senior care facility cluster. Three-quarters of the Hokkaido cancer center cluster were female, while the female-to-male ratio for the gyms and university-related parties cluster were nearly evenly split

between males and females. The proportion of cases with no reported date of onset ranged from 2.5% for the Aichi gyms cluster to 46.5% for the Kyoto university-related parties cluster. The time between first onset and last onset within each cluster ranged from 22 to 43 days.

The delay from onset to prefecture report date for cases reported between when the first case was detected in January and May 31, 2020 was estimated at 7.2 days (95% credible interval [CrI]: 7.1–7.3 days) using the best-fit gamma distribution (see Table S1 and Figure S1). Figure 1 depicts the probability distributions of observing additional cases by cluster, varyingly accounting for asymptomatic cases and missing dates of onset (imputed dataset) and underascertainment of cases.

Table 3. Parameters used in the statistical model.

Description	Distribution	Examined values	References
Offspring distribution	Negative binomial	R_e: 0.5, 1.5, 3.0 k (SD): 0.11 (0.05), 0.25 (0.19), 0.58 (0.26)	Park (Park et al., 2020a) and Tariq, Zhang, Bi (Bi et al., 2020; Tariq et al., 2020; Zhang et al., 2020b)
Serial interval	Weibull	Mean: 4.8 (SD: 2.3)	Nishiura (Nishiura et al., 2020e)
Reporting delay	Gamma	Mean: 7.2 (SD: 4.7)	Estimated

R_e: effective reproduction number. k: overdispersion parameter. SD: standard deviation.

Table 4. Descriptive information regarding the four case clusters.

Characteristics	Fitness gyms	University parties	Senior care	Cancer center
Source of index case infection	Imported	Imported	Domestic	Domestic
Number of cases	40	72	94	92
Age in years,* mean (range)	59 (20–80)	30 (0–70)	72 (20–100)	50 (20–80)
Female (%)	19 (47.5)	33 (45.8)	71 (75.5)	62 (67.4)
Date of onset not reported (%)	1 (2.5)	33 (45.8)	34 (36.2)	24 (26.1)
First to last onset	35 days	22 days	35 days	43 days

*Ages were reported in deciles. Source of index case infection "imported" indicates that the index case(s) for the cluster reported international travel and were likely infected while abroad.

Table 5 presents the estimated observation period in days for each cluster based on the varying parameter values. End-of-outbreak determination was most closely tied with the size of the outbreak. The datasets with more cases (accounting for missing onsets and underascertainment of cases) reached the proscribed probability thresholds at later dates compared to the datasets based on the original epidemic curve (the reported dataset). Studies on Ebola virus disease likewise found that lower ascertainment of cases indicates that more time is needed to be sufficiently certain that the end of an outbreak is declared appropriately (Lee and Nishiura, 2019; Thompson et al., 2019). As well, a larger R_e consistently resulted in slightly longer observation periods compared to smaller R_e (see Table 3 and Figures S2.1– 2.4). The CrI were widest for the combination of a large R_e and small k, indicating greater uncertainty in whether the outbreak had truly ended. The probabilities for some of the upper 95% CrI never dropped below the threshold values within the 42-day periods examined.

Table 5. Length of the observation period in days from last date of report of a case to estimated end of transmission for four coronavirus disease 2019 (COVID-19) case clusters in Japan

Cluster			A	B	C	D
Parameters	R_e	k	Probability threshold (5%)			
Reported onsets	0.5	0.11	23	13	22	10
		0.25	23	13	22	10
		0.58	23	13	22	10
	1.5	0.11	27	17	26	14
		0.25	27	17	26	14
		0.58	27	17	26	14
	3.0	0.11	29	20	29	17
		0.25	29	20	28	17
		0.58	29	20	28	17
Imputed onsets	0.5	0.11	23	14	22	17
		0.25	23	14	22	17
		0.58	23	14	22	17
	1.5	0.11	27	18	26	21
		0.25	27	18	26	21
		0.58	27	18	26	22
	3.0	0.11	30	21	29	24
		0.25	29	21	29	24
		0.58	29	21	29	24
20% under-ascertainment	0.5	0.11	30	22	30	28
		0.25	30	22	30	28
		0.58	30	22	31	28
	1.5	0.11	35	27	35	33
		0.25	35	26	35	32
		0.58	34	26	35	32
	3.0	0.11	-	-	-	-
		0.25	38	29	38	35
		0.58	37	29	37	34

Cluster A: Fitness gyms cluster; Cluster B: University parties cluster; Cluster C: Senior care facility; Cluster D: Cancer center cluster. Observation days are reported relative to the last date of report of a case in the cluster (Day 0). Cells with "-" did not reach the proscribed threshold probability within the 42-day period analyzed.

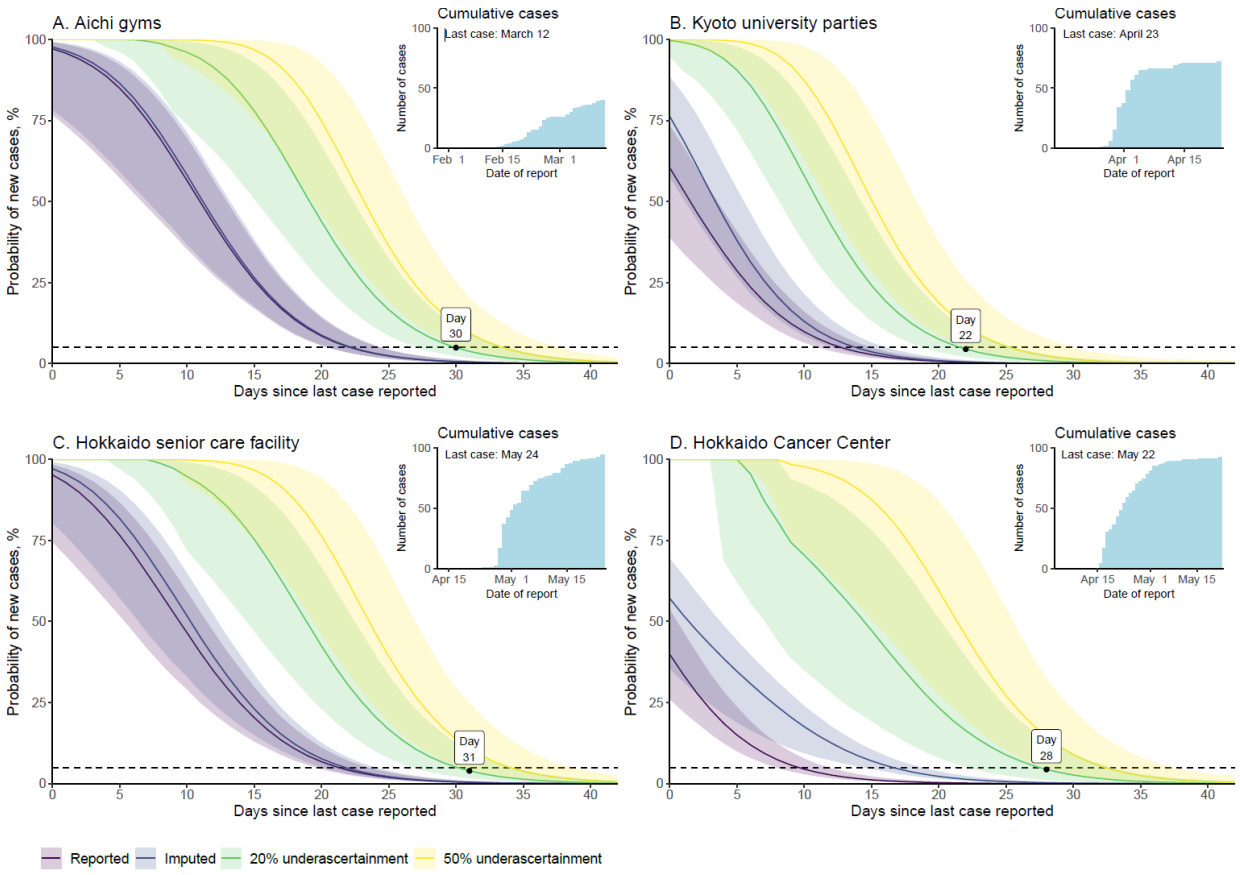

Figure 3. End-of-outbreak probabilities for four coronavirus disease 2019 (COVID-19) case clusters in Japan.

Each subfigure begins at the last date of onset within the cluster. All plots assume $R_e = 0.5$ and k=0.25. Lines are median values and shaded areas are 95% confidence intervals (CrI) for the datasets. Purple represents the datasets including only reported dates of onset; indigo represents the datasets including imputed dates of onset; yellow represents the datasets accounting for 20% underascertainment of cases; green represents the datasets accounting for 50% underascertainment of cases. The horizontal line represents the threshold for 5% probability of failure of the model. Cumulative case counts over time for each cluster are shown in the inset figures.

Discussion

As guidance continues to be developed for COVID-19 response, it is important to incorporate insights from statistical modeling to declining phases of the pandemic (Thompson et al., 2020). In this regard, localized end-of-outbreak declarations are a valuable component of outbreak response as they indicate that public health interventions may be relaxed or lapsed. Statistical models—which can provide insight into levels of certainty of the decline of an outbreak—are useful for dynamically optimizing the timing of such declarations.

Our application of a transmission characteristic-based statistical model to COVID-19 clusters found that confidence in end-of-outbreak determination was most closely tied to estimates of R_e and k, which in turn reflect local levels of control. If public health measures are effective, we would expect to see lower R_e (less transmission on average) and larger k (lower risk of superspreading due to measures targeted to prevent superspreading events) and can have more confidence in declaring the end of the outbreak. When considering end-of-outbreak scenarios we would generally anticipate $R_e < 1$. However, if no or lax public health interventions are implemented R_e may be > 1. Under these circumstances, there is less certainty as the end-of-outbreak probabilities approach zero (see supplementary figures). Under some assumptions of R_e and k, we estimated that end-of-outbreak declarations could be made before 28 days—two times the approximate maximum incubation period for COVID-19 (Linton et al., 2020; Public Health Agency of Canada, 2020)—had passed from the date of last report of any case (which is often when the same day the case is isolated), and using these estimates could potentially allow for earlier end-of outbreak declaration, leading to saved resources.

Parameters such as the reporting delay and serial interval may also vary throughout the epidemic. When surveillance is heightened the reporting delay may be shorter than when surveillance systems are overwhelmed. Similarly, nonpharmaceutical interventions such as contact tracing, isolation, and physical distancing change contact patterns and limit the time during which an infectious case may be in contact with susceptible individuals, shortening the serial interval (Ali et al., 2020). Although these possible variations were not accounted for in this study, they can be incorporated if deemed to be of value to inform decisions regarding the continuation of public health and social response measures.

Furthermore, the size and scale of the epidemic curves used in our analyses depend on case, cluster, and outbreak definitions and case ascertainment by the surveillance system. When a cluster definition is limited to cases directly linked to a location or activity (e.g., a hospital or an event) then the cluster size will be smaller than if the cluster includes all secondary infections to household members and other contacts not directly related to the transmission event(s) that defined the cluster. Likewise, when the cases are limited to those whose samples test positive (i.e. via polymerase chain reaction [PCR] or antigen testing) the scope and scale of the epidemic curve will be smaller than if other probable cases were included in the outbreak case definition. The clusters reported here include original cluster cases as well as all subsequent cases in chains of transmission reported by local public health jurisdictions. It is possible that some cases associated with the cluster were missed or incorrectly attributed; however, we have repeatedly reviewed the data to minimize these possibilities. Our analyses accounting for possible underascertainment of cases likewise show that when accounting for missed cases extends the observation period before reaching our 5% threshold is extended.

Through the end of May 2020, only PCR-positive cases were included in the case definition for COVID-19 cases in Japan. Infected individuals may not have been tested if they were never suspected of being a case or did not meet testing criteria (Ministry of Health Labour and Welfare (MHLW)). In addition, PCR sensitivity is less than perfect, reducing to around 70% more than one week after symptom onset (Miller et al., 2020), so some infected individuals may have received a false negative test result if their viral load at the time they were tested was insufficient to trigger a positive result. Sequentially repeated PCR testing in Japan for persons with persistent symptoms and/or new onset of symptoms after initially being tested while asymptomatic has identified cases that were initially PCR negative but epidemiologically linked to other cases, as has been seen elsewhere (Ai et al., 2020).

In addition, importation of cases is not accounted for in this method. Defining outbreaks based on the epidemiological linkage of cases to at least one common source of exposure (i.e. clusters) necessarily precludes inclusion of new sources of infection (i.e. importation). A new case linked to—for example—a physical location that was a common source of exposure for cases in a cluster may represent an importation event rather than a continuation of the outbreak/cluster unless there is clear epidemiological link (e.g., close contact or physical proximity during a given

timeframe) between the newly detected case and the cluster. When applying this method to outbreaks defined by a geographic region with free-flow borders to other regions with active transmission—as was the cases for prefectures in Japan during the first wave of the pandemic—importation of one or more cases before the outbreak would simply add to the existing epidemic curve and funnel into the end-of-outbreak probability calculations. Likewise, exportation of cases is not accounted for in this method, as any case epidemiologically linked to the cluster is included, regardless of geographic boundaries within Japan. However, possible exportation across international borders is not accounted for, and even when examining clusters if local public health jurisdictions minimize publicly shared information (as was seen in later stages of the pandemic), some links may be missed.

Lastly, further analyses describing transmissivity of the virus are needed to improve understanding of the most plausible range for these critical values. In addition, when more information is available regarding the different transmission routes of COVID-19 (i.e. airborne, droplet spread or contact with contaminated fomites) the model could potentially be updated to account for differences in these routes, as it was previously done for modeling the flare-ups of Ebola virus disease due to sexual transmission from male survivors (Lee and Nishiura, 2019). Other methods for statistical end-of-outbreak determination have recently been proposed, and provide alternative options for examining cases using geographically-based outbreak definitions (Hart et al., 2019; Parag et al., 2020). The focus of this study on clusters may be different from typical geographically-based analyses, but we believe focusing on this scale can be meaningful to decision-makers dealing with clusters on an individual basis, such as officials for the involved local health jurisdictions (i.e. cities and prefectures) as well as the facilities (hospitals, senior homes, gyms, schools, etc.) where cases have been identified.

In summary, we incorporated use of the reporting delay distribution into a model for end-of-outbreak probability estimation and applied this method to clusters in the COVID-19 epidemic in Japan. In doing so, we provide estimates of the probability that the outbreak will continue in real time. Communicating these probabilities can inform public health decision-making around the appropriate use of resources when transmission has declined for a given outbreak.

Chapter 2: Supplementary materials

Clusters

The Hokkaido cancer center cluster excludes two reports of cases who re-tested positive after being considered recovered.

Serial interval

The mean μ and shape parameter k for the Weibull distribution were drawn from normal distributions with mean and standard deviation (SD) for these values as reported by Nishiura et al. (Nishiura et al., 2020e). The scale λ parameter was calculated as:

$$\lambda = \frac{\mu}{\Gamma(1 + 1/k)} \tag{S6}$$

Reporting delay and end-of-outbreak probability estimates

Estimates of the reporting delay are shown in Table S1. The empirical reporting delay is shown in Figure S3. Figures S4-7 show a sensitivity analysis of the estimated end-of-outbreak probabilities for different values of R_e and k.

Table S1. Comparison of fitted distributions for the reporting delay of COVID-19 cases confirmed in Japan, February–May 2020.

Distribution	Mean, days	Standard deviation, days	LOOIC
Gamma	7.2 (7.1–7.2)	4.7 (4.6–4.8)	52577
Weibull	7.2 (7.1–7.2)	4.6 (4.5–4.6)	52719
Lognormal	7.5 (7.4–7.6)	6.1 (5.9–6.3)	53273

LOOIC: leave-one-out information criteria calculated according using loo 2.2.0. 95% confidence intervals are shown in parenthesis.

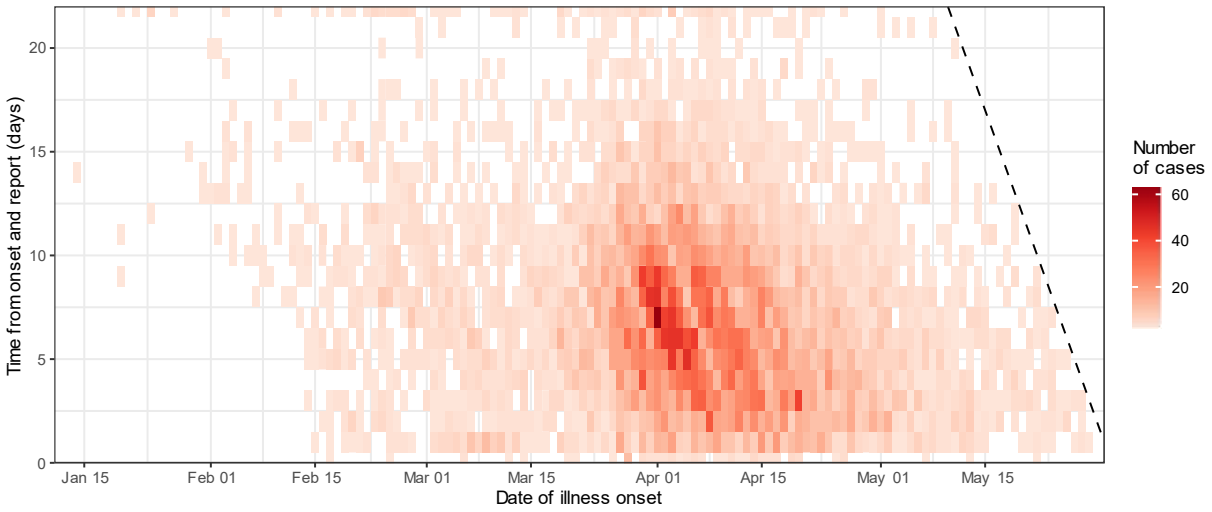

Figure S3. Reporting delay for COVID-19 cases confirmed in Japan, January–May 2020. Estimates are from 6,921 cases with onset available reported through May 31, 2020. Excludes cruise ship cases, evacuees from Wuhan, and international travelers. Darker red indicates a greater number of cases. There are no observations to the right of the dashed line due to truncating the data to exclude cases reported after May 31.

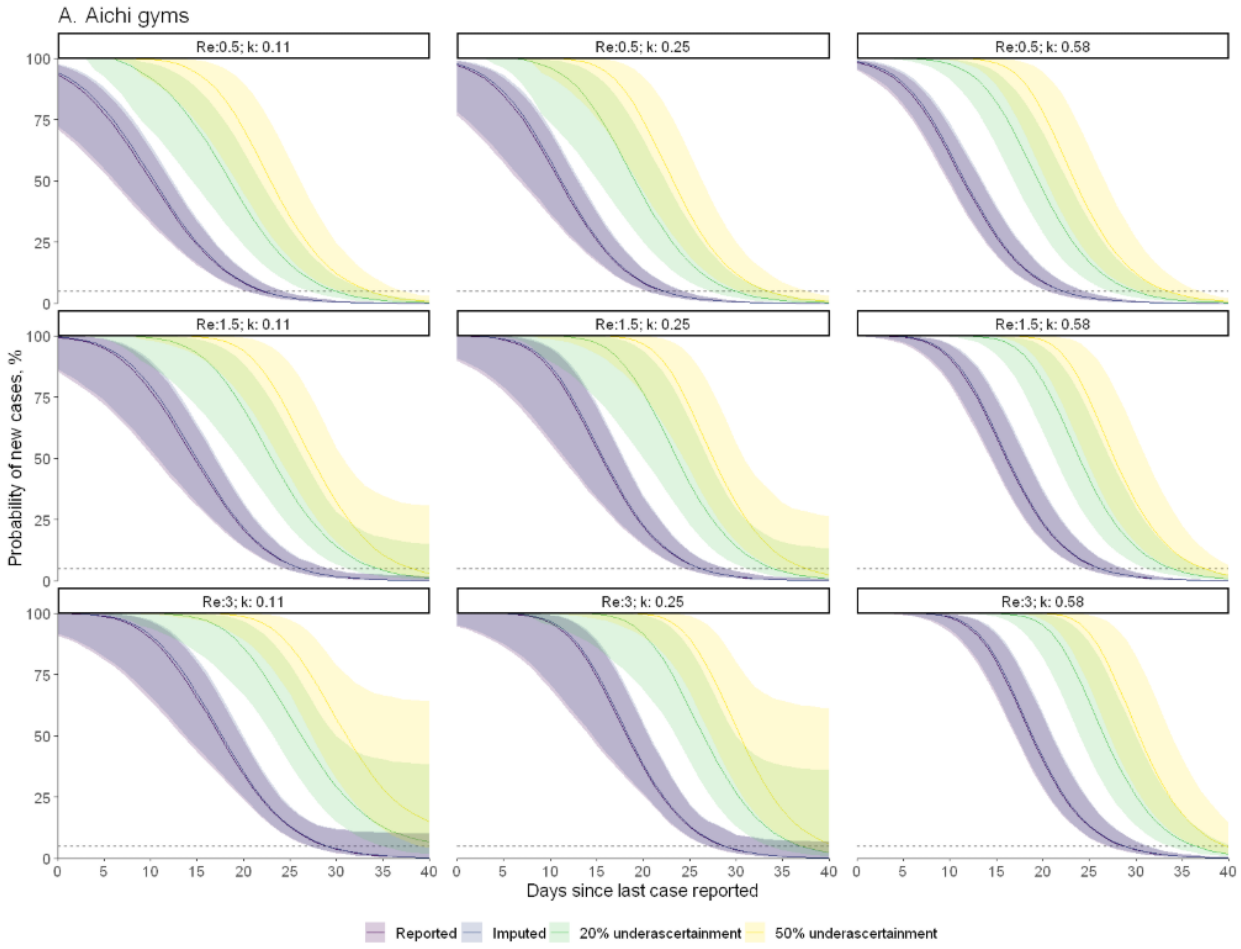

Figure S4. End-of-outbreak probabilities for the Aichi fitness gyms cluster. Purple represents the datasets including only reported dates of onset; indigo represents the datasets including imputed dates of onset; yellow represents the datasets accounting for 20% underascertainment of cases; green represents the datasets accounting for 50% underascertainment of cases. The horizontal line represents the threshold for 5% probability of failure of the model.

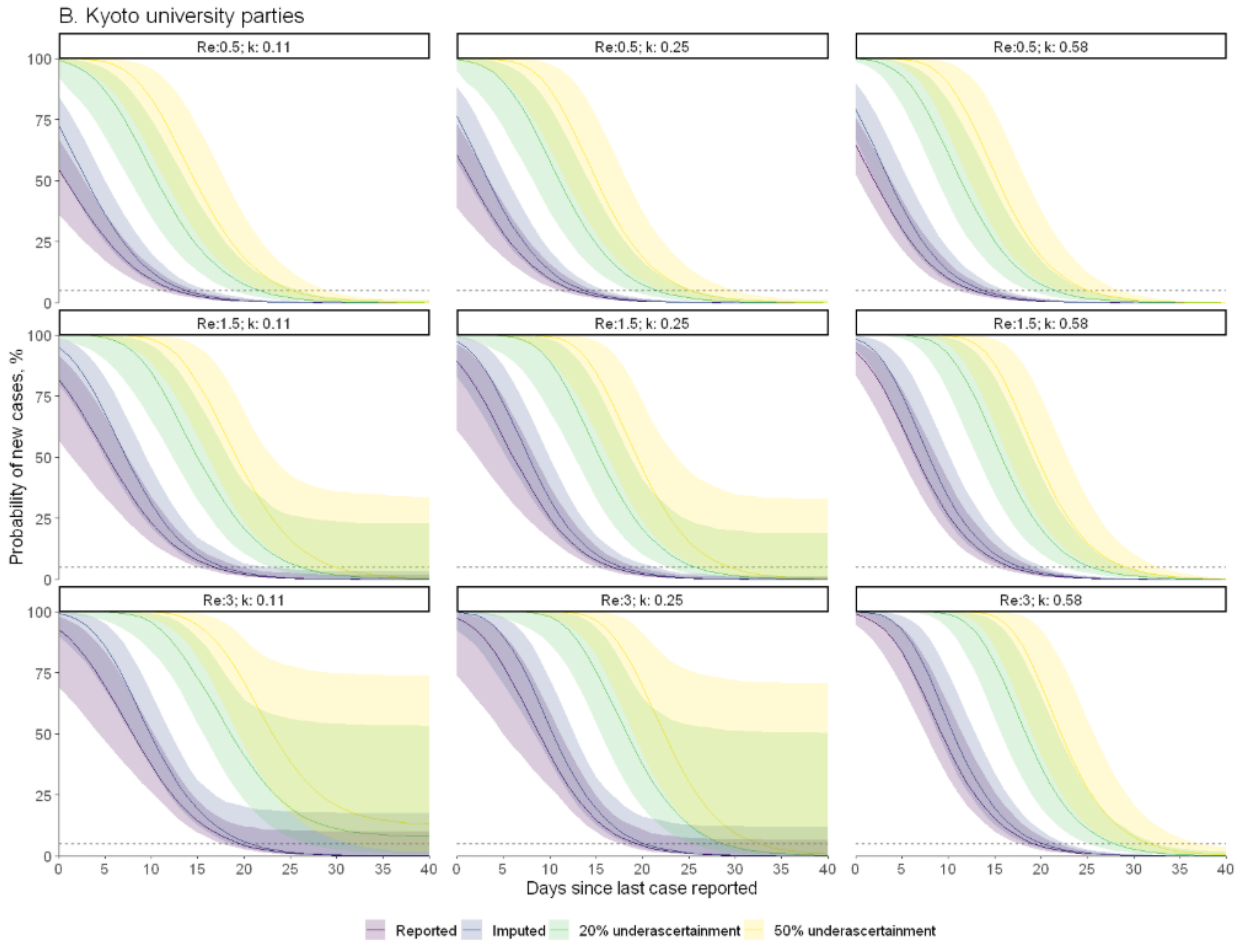

Figure S5. End-of-outbreak probabilities for the Kyoto university parties cluster. Purple represents the datasets including only reported dates of onset; indigo represents the datasets including imputed dates of onset; yellow represents the datasets accounting for 20% underascertainment of cases; green represents the datasets accounting for 50% underascertainment of cases. The horizontal line represents the threshold for 5% probability of failure of the model.

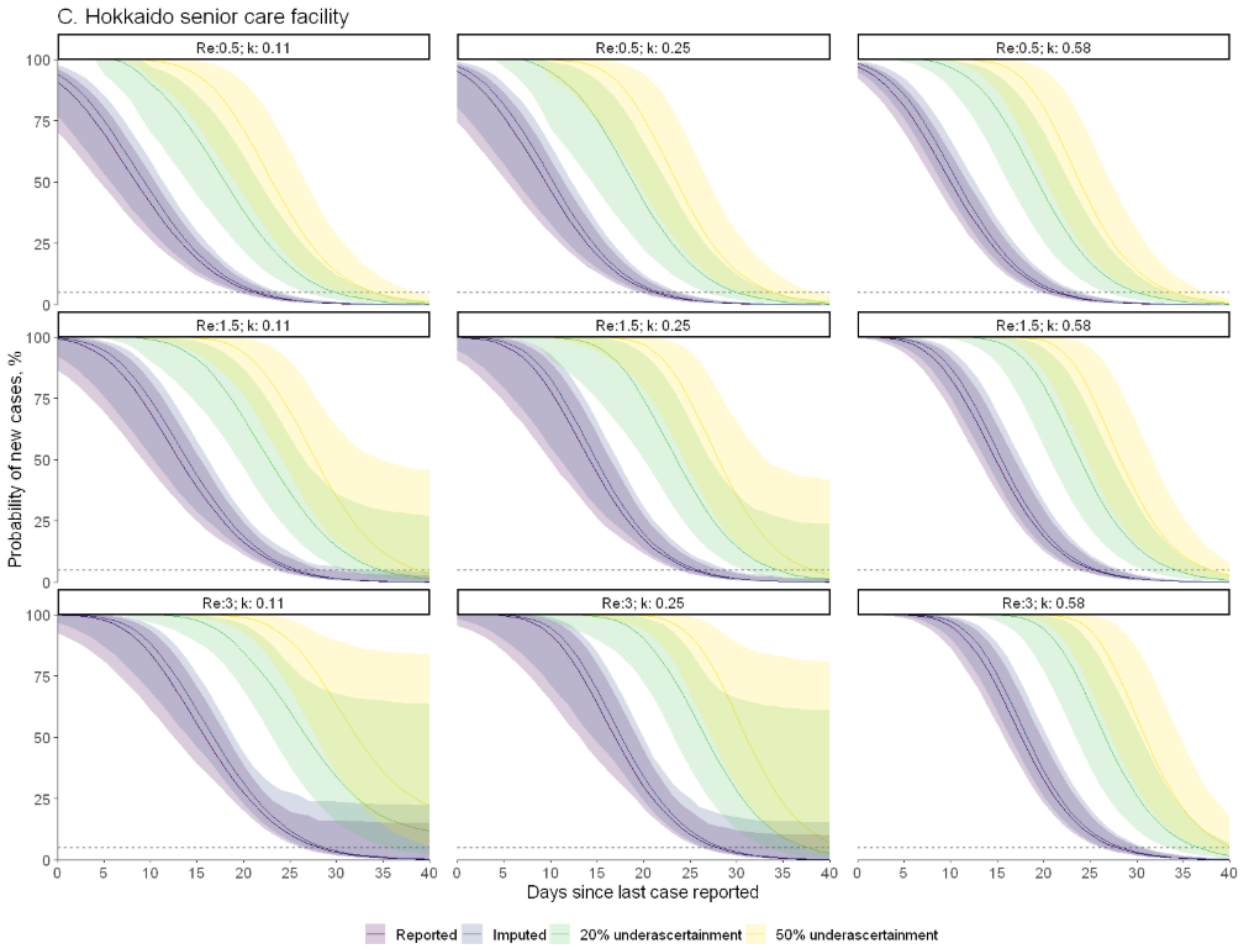

Figure S6. End-of-outbreak probabilities for the Hokkaido senior care facility cluster. Purple represents the datasets including only reported dates of onset; indigo represents the datasets including imputed dates of onset; yellow represents the datasets accounting for 20% underascertainment of cases; green represents the datasets accounting for 50% underascertainment of cases. The horizontal line represents the threshold for 5% probability of failure of the model.

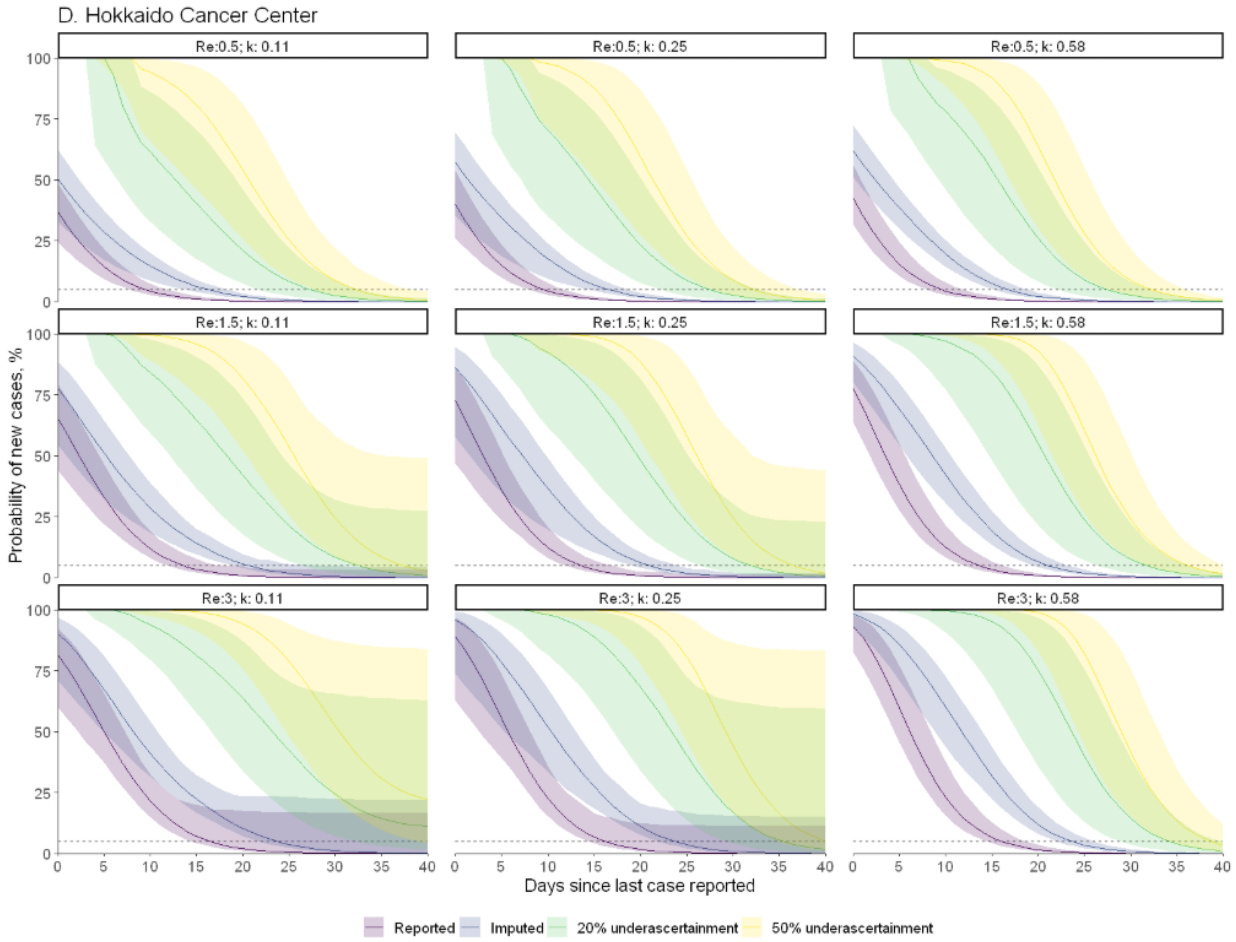

Figure S7. End-of-outbreak probabilities for the Hokkaido cancer center cluster. Purple represents the datasets including only reported dates of onset; indigo represents the datasets including imputed dates of onset; yellow represents the datasets accounting for 20% underascertainment of cases; green represents the datasets accounting for 50% underascertainment of cases. The horizontal line represents the threshold for 5% probability of failure of the model.

Chapter 2 was originally published as:

Linton NM, Akhmetzhanov AR, Nishiura H. Localized end-of-outbreak determination for coronavirus disease 2019 (COVID-19): examples from clusters in Japan. *Int J Infect Dis*. 2021;105:286-292. Minor formatting modifications and edits have been made for the dissertation.

Chapter 3: Correlation between times to SARS-CoV-2 symptom onset and secondary transmission undermines epidemic control efforts

Background

The generation interval and incubation period of an infectious disease are key epidemiological parameters used to inform outbreak response. The former describes the infectiousness of a pathogen in a host—it represents the time between when a host (an infector) is infected with a pathogen and when they transmit that pathogen to another host (an infectee). In contrast, the incubation period represents the time between infection and development of symptomatic disease. It reflects the pathogen replication rate and provides a basis for predicting prognosis. It also indicates how long an infection may remain unnoticed in an individual, and thus for emerging diseases it is used to determine quarantine periods (Linton et al., 2020). Together, the generation interval and incubation period provide insight into the dynamics of an infectious disease and characterize the effectiveness of public health interventions on its control.

At the beginning of the COVID-19 pandemic, the mean incubation period of SARS-CoV-2—the pathogen causing COVID-19—was rapidly estimated (Backer et al., 2020; Lauer et al., 2020; Li et al., 2020; Linton et al., 2020). Estimates of the mean serial interval, which is the time between symptom onset in an infector and symptom onset in a person they infect, quickly followed (Ferretti et al., 2020; Nishiura et al., 2020e; Tindale et al., 2020). However, due to difficulty in ascertaining exposure times of cases and directionality of transmission between epidemiologically linked cases, few attempts were made to estimate the generation interval (Deng et al., 2020; Ferretti et al., 2020; Ganyani et al., 2020; Li et al., 2021; Tindale et al., 2020). Instead, the serial interval was often used as a proxy for the generation interval when estimating epidemiological quantities. For example, it has been used to estimate the basic reproduction number—the average number of persons infected by a single infector in a completely susceptible population—and the effective reproduction number—the average number of persons infected by a single infector in the presence of existing infections and interventions (Abbott et al., 2020; Jung et al., 2020c). However, use of the serial interval as a proxy for generation time can lead to biased estimates of the effective reproduction number due to factors such as differences in their variances and the ability for the serial interval to have negative values—i.e. infectee onset

preceding infector onset (Deng et al., 2020; Knight and Mishra, 2020; Park et al., 2021; Torneri et al., 2021).

Generation intervals depend on many factors, such as the ratio of infected to susceptible persons among contacts, as well as the behavior of infected persons (Nishiura, 2010). Actions such as self-isolation after symptom onset can shorten generation intervals by limiting the opportunity for infected individuals to infect others. However, if an infected person can transmit the pathogen before symptom onset, then isolation of symptomatic persons alone is insufficient to control the spread (Fraser et al., 2004). For pathogens such as SARS-CoV-2, interventions targeting nonsymptomatic cases appears vital (Lovell-Read et al., 2021) due to the large fraction of presymptomatic and asymptomatic transmission (He et al., 2020a; Nakajo and Nishiura, 2021; Tindale et al., 2020).

Frequently, estimates of the generation interval of SARS-CoV-2 have been derived from the serial interval and formulated following the assumptions that: i) there is no asymptomatic transmission (Inaba and Nishiura, 2008), ii) the incubation period and generation interval are independent (Lehtinen et al., 2021). However, these assumptions are clearly flawed with respect to SARS-CoV-2 transmission. In the case of asymptomatic transmission, there is substantial evidence of transmission from asymptomatic infectors (Gandhi et al., 2020; Nakajo and Nishiura, 2021), as well as plenty of asymptomatic cases demonstrating the existence of asymptomatic infectees. Likewise, correlation between the generation interval and incubation period of SARS-CoV-2 was shown to be biologically plausible, as evidenced by the peak viral load of SARS-CoV-2 occurring around the time of symptom onset (He et al., 2020b). However, previous studies that considered such a correlation did not attempt to directly estimate it (Bushman et al., 2021). Here, we provide direct evidence of their correlation in this study.

In this study, we jointly estimated the generation interval and incubation period of COVID-19 using transmission pairs identified in Japan in 2020 and quantified their dependence. We assessed whether either interval or their correlation varied based on sociodemographic, epidemic, or transmission characteristics. Accounting for correlation between the generation interval and incubation period can improve estimation of SARS-CoV-2 transmission as well as the effectiveness of public health interventions by preventing underestimation of the proportion of

presymptomatic transmission and the effect of isolation of symptomatic cases on epidemic control.

Methods

Study population and setting

We compiled a dataset of COVID-19 transmission pairs using openly published case data from reporting jurisdictions (prefectures and cities) in Japan, focusing on detecting pairs for whom directionality of transmission could be determined with some degree of certainty. Cases were limited to those reported during the calendar year 2020. Jurisdiction reporting practices changed over time, with details generally becoming sparser over time, as concerns grew around infection-related stigmatization (Yoshioka and Maeda, 2020), as well as in prefectures with large case loads.

Among the information publicly shared were links between cases and links to common exposures (e.g., a medical facility, event, or restaurant). However, clear statements as to who was the infector between linked cases or within clusters were generally not published. As well, dates of contact between cases and details of the type of link between cases were often only reported in detail if deemed to be important for public health action, limiting the number of cases for whom detailed epidemiological information related to their linkages were available. Therefore, assumptions about directionality of transmission were largely at the discretion of the authors, and in consequence we used the following bases for identifying linked cases as directional transmission pairs: 1) linkage of the infector (but not infectee) to a cluster; 2) the dates of contact, type of contact, and onset dates reported for linked cases provided some insight into directionality of transmission; 3) the infector or index case of a chain travelled to a location with increased/increasing transmission prior to onset; or 4) the infector or index case of a chain was presumed to have been infected while travelling abroad. Households with >2 cases and links between cases in clusters where directionality of transmission and timings of contact could not be clearly identified were not selected (Britton and Tomba, 2019).

We included pairs with infectors who had multiple possible exposures (their exposure period takes the lower and upper bounds of all possible exposures) but excluded possible pairs where potential infectees had multiple possible infectors, and it is possible that infectees with multiple potential infectors would have different contact patterns (and possibly be associated with shorter generation intervals) compared to infectees that had only one potential infector identified, as a susceptible person is likely to become infected more quickly if they are surrounded by multiple possible sources of infection (Kenah et al., 2008; Nishiura, 2010). Further details regarding ascertainment of transmission pairs are available in the supplementary information.

Exposures were defined in relation to travel, contact with a confirmed case, or link to a cluster/common exposure. Reports of symptom onset in Japan were not restricted to any particular symptom, such as fever, but may have been reported as beginning with any of a variety of symptoms associated with SARS-CoV-2 infection such as fever, cough, fatigue, or runny nose.

Data stratification

The dataset including coarsely reported dates of exposure and contact were divided into subgroups to assess whether the generation interval, incubation period, or correlation between the two parameters would vary by subpopulation. Age (reported in deciles) was divided into three groups: cases under 30 years of age, cases 30–59 years of age, and cases 60+ years of age. Sex was reported as female or male. Separate age and sex subgroups were established for infectors and infectees. Type of contact between infector and infectee was divided into three categories: household contact, social contact-based interaction, and core community interaction. These divisions were made with public health interventions in mind. For example, social contact-based interaction includes types of contact that may not have occurred when local control measures were advised or a state of emergency was declared (Cabinet Secretariat of Japan, 2021).

Generally, public health control measures in Japan promoted during 2020 focused on reducing the number of people individuals were physically in contact with in a given day, as well as reducing scenarios where the “Three C’s”— closed spaces, crowded places, and close-contact settings—were present (Oshitani and The Expert Members of the National COVID-19 Cluster Taskforce at the Ministry of Health Labour and Welfare of Japan, 2020). Interventions in Japan

included limiting the total number or proportion of people who can visit facilities and venues, limiting restaurant hours, encouraging staying at home and discouraging cross-prefecture travel, etc. Our definition of ore community interaction, in contrast, focuses more on contact that occurs in schools, workplaces for general business, essential workplaces (medical facilities, care facilities, government services, etc.), and unknown sources of infection (community infection). Although these settings assigned to the core community interaction category may also be targeted by public health measures, they are perhaps less acutely affected by government decrees and social sentiment compared to settings more closely related to social contact-based interaction (Cabinet Secretariat of Japan, 2021).

Japan experienced three waves of COVID-19 during 2020, with the third wave extending into 2021 (Figure S9). The first wave began with the first reported case, confirmed to be positive for SARS-CoV-2 on 16 January 2020. The second wave we set to begin on 1 June, which is around the center of the bottom of the trough between the peaks of the first and second waves. The third wave we set to begin on 1 October, which likewise is around the center of the bottom of the trough between the peaks of the second and third waves. Assignment to a given wave for each pair was determined by infector report date. Lastly, to check for differences given our basis for selecting transmission pairs we also stratified the dataset according to whether directionality was determined with respect to 1) importation from abroad, 2) linkage of the infector (but not infectee) to a cluster, 3) domestic travel by the infector to a location with increasing transmission, or 4) the timing and type of contact between cases in transmission chains.

Statistical analyses

Descriptive analyses and visualization were performed using R 4.1.0 (R Development Core Team, 2019). Bayesian parameter estimation was implemented in Stan using the cmdstanr interface to CmdStan 2.26.1 (Stan Development Team, 2021). To assess correlation between the generation interval and incubation period of the infector we constructed a joint probability distribution for the generation interval and infector incubation period by use of copulas—multivariate cumulative distribution functions (Klinkenberg and Nishiura, 2011; Sklar, 1959). The copulas we assessed included the Gaussian (normal), Clayton, Gumbel, and independence copulas. They are described in detail in the Supplementary Materials. Timing of pathogen

transmission and symptom onset was estimated using interval censoring methods derived from Reich et al. (2009) and adapted from previously published work (Linton et al., 2020; Nishiura et al., 2020e). For all parameters, posterior point estimates are given by the 50th percentiles of the converged Markov chain Monte Carlo (MCMC) chains, and the best combination of copula and parametric distributions were selected using weights from a Bayesian mixture model (see Supplementary Materials).

To consider the effect of correlation between the generation interval and incubation period on transmission, we simulated 10,000 pairs from the best-fit model of the jointly estimated generation interval and incubation period for all possible values of Kendall's tau from 0 to 1. For each simulated pair, we determined whether transmission was presymptomatic based on whether the incubation period was greater than the generation interval, and thereby calculated the proportion of presymptomatic transmission p for the 10,000 pairs for each value of Kendall's tau. We then considered that symptomatic transmission could result in a decrease in transmission as defined by the basic reproduction number R_0—the average number of infectees generated by a single infector. We calculated the effective R_0 as $R_0^E = \varphi R_0 + (1-\varphi)(1-\varepsilon)R_0$ where φ is the proportion of presymptomatic transmission and ε is the percent reduction in transmission due to rapid isolation among symptomatic cases, considering $R_0 = 2.2$ (Riou and Althaus, 2020) and varied this by ±0.7, also considering $R_0 = 1.5$ and $R_0 = 2.9$.

Results

Characteristics of transmission pairs

Information on timing of exposure and onset for infectors as well as contact between infectors and infectees was obtained for 286 transmission pairs reported in Japan during 2020, of which 257 pairs had symptom onset available for the infector. For the other 29 pairs the infectors were asymptomatic at time of report. Of the 257 pairs with symptomatic infectors, 49 (19.2%) had single dates reported for both infector exposure and contact between infector and infectee. Characteristics of the pairs in each dataset are shown in Table 1, while Figure S10 provides insight into the relationship between the empirical generation intervals, serial intervals, and incubation periods associated with these cases.

For the dataset of 257 pairs with symptomatic infectors, most infectors (50.2%) and infectees (44.4%) were between 30–59 years of age. There were fewer female infectors (30.7%) detected compared to female infectees (51.0%). Age and sex distributions of infectors and infectees are shown in Figure S11. Pairs were relatively evenly distributed between the three pandemic waves that occurred during 2020. Most pairs (53.27) were linked to a cluster, while other pairs were identified by having contact patterns indicative of directionality of transmission (21.0%), the infector had travel to another prefecture before onset (16.7%), or the infector was an imported case or otherwise linked to an imported case (8.6%). Given that it was easier to determine the directionality of pairs and obtain information on timing of exposure if the infector was linked to a cluster—an aggregation of cases with a common exposure—or had travel history, our dataset includes only a handful of infectors (1.6%) with household exposure. In contrast, nearly half (42.4%) of infectees were household/family members of their infectors. The single-date (49 pairs) and asymptomatic infector (28 pairs) datasets were similarly structured in terms of age and sex, though only one asymptomatic infector was detected during the first wave.

Table 6. Characteristics of transmission pairs

Subgroups	Status	Pair characteristics	Coarse dates of exposure† (%)	Single dates of exposure‡ (%)	Asymptomatic infectors (%)
		All	257 pairs	49 pairs	29 pairs
Age	Infector	Under 30 years	65 (25.3%)	11 (22.4%)	6 (20.7%)
		30-59 years	129 (50.2%)	32 (65.3%)	15 (51.7%)
		60+ years	57 (22.2%)	6 (12.2%)	7 (24.1%)
		Not reported	6 (2.3%)	-	1 (3.4%)
	Infectee	Under 30 years	69 (26.8%)	17 (34.7%)	9 (31.0%)
		30-59 years	114 (44.4%)	19 (38.8%)	10 (34.5%)
		60+ years	67 (26.1%)	13 (26.5%)	9 (31.0%)
		Not reported	7 (2.7%)	-	1 (3.4%)
Sex	Infector	Female	79 (30.7%)	17 (34.7%)	12 (41.4%)
		Male	177 (68.9%)	32 (65.3%)	16 (55.2%)
		Not reported	1 (0.4%)	-	1 (3.4%)
	Infectee	Female	131 (51.0%)	22 (44.9%)	16 (55.2%)
		Male	125 (48.6%)	27 (55.1%)	12 (41.4%)
		Not reported	1 (0.4%)	-	1 (3.4%)
Transmission setting	Infector	Household	4 (1.6%)	0 (0.0%)	0 (0.0%)
		Social	153 (59.5%)	48 (98.0%)	25 (86.2%)
		Community	100 (38.9%)	1 (2.0%)	4 (13.8%)
	Infectee	Household	109 (42.4%)	2 (4.1%)	11 (37.9%)
		Social interaction	113 (44.0%)	38 (77.6%)	15 (51.7%)
		Community	35 (13.6%)	9 (18.4%)	3 (10.3%)
Epidemic wave		Wave 1	86 (33.5%)	14 (28.6%)	2 (6.9%)
		Wave 2	95 (37.0%)	16 (32.7%)	13 (44.8%)
		Wave 3	76 (29.6%)	19 (38.8%)	14 (48.3%)
Basis for selection		Cluster	138 (53.7%)	29 (59.2%)	24 (82.8%)
		Contact pattern	54 (21.0%)	19 (38.8%)	4 (13.8%)
		Domestic travel	43 (16.7%)	1 (2.0%)	1 (3.4%)
		Import	22 (8.6%)	-	-
Region§		Hokkaido & Tohoku	35 (13.6%)	6 (12.2%)	2 (6.9%)
		Kanto	31 (12.1%)	4 (8.2%)	0 (0.0%)
		Chubu	73 (28.4%)	12 (24.5%)	11 (37.9%)
		Kinki	38 (14.8%)	6 (12.2%)	3 (10.3%)
		Chugoku & Shikoku	34 (13.2%)	6 (12.2%)	6 (20.7%)
		Kyushu & Okinawa	46 (17.9%)	15 (30.6%)	7 (24.1%)

†Dataset using intervals of exposure and/or contact between infector and infectee. ‡Dataset limited to pairs where infector exposure and infector-infectee contact were limited to a single day. “Hokkaido & Tohoku” includes Hokkaido, Aomori, Iwate, Miyagi, Akita, Yamagata, and Fukushima prefectures. “Kanto” includes Ibaraki, Tochigi, Gunma, Saitama, Chiba, Tokyo, and Kanagawa prefectures. “Chubu” includes Niigata, Toyama, Ishikawa, Fukui, Yamanashi, Nagano, Gifu, Shizuoka, and Aichi prefectures. “Kinki” includes Mie, Shiga, Kyoto, Osaka, Hyogo, Nara, and Wakayama prefectures. “Chugoku & Shikoku” include Tottori, Shimane, Hiroshima, Yamaguchi, Tokushima, Kagawa, Ehime, and Kochi prefectures. “Kyushu & Okinawa” include Fukuoka, Saga, Nagasaki, Kumamoto, Oita, Miyazaki, Kagoshima, and Okinawa prefectures.

Joint estimates of the generation interval and incubation period

The mean generation interval ranged between 3.7 and 5.1 days, with the mean for the overall dataset estimated at 4.3 days (95% CrI: 4.0–4.7 days). In contrast, the estimated generation interval for the dataset of asymptomatic infectors was longer, at 4.6 days (3.9–5.5 days), resulting in a ratio of asymptomatic-to-symptomatic generation intervals of 1.1. The jointly estimated mean incubation period was consistently longer than the generation interval, ranging from 4.4–5.7 days, and estimated at 4.7 days (95% CrI: 4.3–5.0 days) for the overall dataset, providing evidence of presymptomatic transmission. The prior and posterior distributions of the generation interval and incubation period are shown in Figure S12. The generation interval and incubation period were positively correlated, with Kendall's tau ranging between 0.4–0.7 and estimated at 0.5 (95% CrI: 0.4–0.6) for the overall dataset (Table 2). For the dataset with single dates of reported exposure and contact, the generation interval was estimated at 4.4 days (95% CrI: 3.9–5.0 days) while the mean incubation period was estimated at 4.9 days (95% CrI: 4.4–5.6 days). Kendall's tau was slightly higher than for the overall dataset, at 0.6 (95% CrI: 0.5–0.7).

The mean generation interval did not vary substantially between subgroups but was shortest for female infectors, at 3.7 days (95% CrI: 3.2–4.4 days). It was also shorter for the second wave of the epidemic (3.8 days, 95% CrI: 3.4–4.3 days) compared to the first wave (5.1 days, 95% CrI: 4.5–5.8 days).27 However, the generation interval for the third wave was longer than that of the second wave—nearly as long as that of the first wave—at 4.5 days (95% CrI: 3.9–5.1 days). Estimates of the incubation period varied less between subgroups, although the mean incubation period for pairs linked to importation from other countries (mostly from the first wave) was a bit longer than the overall estimate, at 5.7 days (95% CrI: 4.6–6.8 days).

The Clayton copula—which emphasizes lower tail dependence—was the most frequently selected copula, although the Gumbel and Gaussian copulas were also selected for some subgroups. The Gumbel copula emphasizes upper tail dependence while the Gaussian copula does not consider tail dependence. The independence copula was never selected (Table 2). For the overall dataset, where the Clayton copula was selected, the lower tail dependence was 0.7 (95% CrI: 0.6–0.8), indicating that infectors with an extremely short incubation period would also be more likely to quickly transmit the virus given contact with a susceptible person (see

Supplementary Materials). For the generation interval, the Weibull distribution was most often selected, although the gamma and lognormal distributions were selected for some subgroups. The lognormal distribution was the only distribution selected across all joint estimates of the incubation period. It is typically the best fit for infectious disease incubation period data (Lessler et al., 2015), including COVID-19 data (McAloon et al., 2020).

Table 7. Joint estimates of the generation interval and incubation period by subgroup

Category	Subgroups	N	Copula	Generation interval			Incubation period			
				Distribution	Mean (95% CrI)	SD (95% CrI)	Distribution	Mean (95% CrI)	SD (95% CrI)	Kendall's tau (95% CrI)
Exact data	All cases	49	Clayton	Weibull	4.38 (3.88, 4.98)	2.10 (1.70, 2.81)	Lognormal	4.92 (4.35, 5.59)	2.66 (2.18, 3.33)	0.61 (0.46, 0.68)
Coarse data	All cases	257	Clayton	Weibull	4.34 (3.98, 4.74)	2.31 (1.99, 2.80)	Lognormal	4.75 (4.42, 5.11)	2.58 (2.25, 2.97)	0.54 (0.44, 0.63)
Infector age	Under 30 years	65	Clayton	Lognormal	4.51 (3.97, 5.14)	2.16 (1.68, 2.90)	Lognormal	4.67 (4.07, 5.37)	2.59 (2.12, 3.27)	0.50 (0.25, 0.65)
	30-59 years	129	Gumbel	Gamma	4.44 (3.97, 4.92)	2.43 (2.01, 3.00)	Lognormal	5.10 (4.63, 5.62)	2.87 (2.45, 3.44)	0.60 (0.48, 0.68)
	60+ years	57	Gumbel	Gamma	4.59 (3.87, 5.35)	2.70 (1.99, 3.70)	Lognormal	4.60 (3.92, 5.35)	2.89 (2.31, 3.69)	0.50 (0.28, 0.65)
Infectee age	Under 30 years	65	Clayton	Lognormal	4.51 (3.97, 5.14)	2.16 (1.68, 2.90)	Lognormal	4.67 (4.07, 5.37)	2.59 (2.12, 3.27)	0.50 (0.25, 0.65)
	30-59 years	114	Gumbel	Weibull	4.65 (4.14, 5.22)	2.59 (2.06, 3.27)	Lognormal	5.26 (4.74, 5.85)	3.12 (2.62, 3.78)	0.58 (0.47, 0.68)
	60+ years	67	Gumbel	Lognormal	4.10 (3.49, 4.84)	2.08 (1.49, 3.00)	Lognormal	4.44 (3.83, 5.15)	2.60 (2.10, 3.29)	0.42 (0.14, 0.61)
Infector sex	Female	79	Gumbel	Gamma	3.72 (3.19, 4.35)	2.09 (1.57, 2.91)	Lognormal	4.38 (3.81, 5.04)	2.59 (2.10, 3.27)	0.54 (0.36, 0.66)
	Male	177	Clayton	Weibull	4.69 (4.27, 5.17)	2.35 (2.00, 2.87)	Lognormal	5.00 (4.58, 5.46)	2.72 (2.34, 3.21)	0.53 (0.40, 0.63)
Infectee sex	Female	131	Clayton	Gamma	4.35 (3.83, 4.96)	2.50 (1.99, 3.30)	Lognormal	4.65 (4.19, 5.16)	2.60 (2.19, 3.16)	0.49 (0.33, 0.63)
	Male	125	Gumbel	Gamma	4.49 (4.07, 4.99)	2.28 (1.89, 2.88)	Lognormal	4.96 (4.49, 5.51)	2.88 (2.44, 3.48)	0.55 (0.43, 0.65)
Epidemic wave	Wave 1	109	Gumbel	Weibull	4.37 (3.74, 4.99)	2.00 (1.59, 2.57)	Lognormal	4.75 (4.21, 5.34)	2.93 (2.45, 3.62)	0.52 (0.31, 0.67)
	Wave 2	113	Clayton	Gamma	4.50 (4.05, 5.04)	2.51 (2.10, 3.14)	Lognormal	4.85 (4.40, 5.38)	2.55 (2.16, 3.07)	0.58 (0.47, 0.67)
	Wave 3	35	Gaussian	Gamma	4.52 (3.83, 5.31)	2.39 (1.74, 3.42)	Lognormal	4.98 (4.19, 5.89)	3.01 (2.38, 3.93)	0.57 (0.37, 0.68)
Transmission setting	Household	86	Gumbel	Weibull	5.08 (4.45, 5.78)	2.81 (2.25, 3.59)	Lognormal	5.12 (4.51, 5.78)	2.87 (2.36, 3.52)	0.55 (0.38, 0.67)
	Social contact	95	Gumbel	Weibull	3.84 (3.42, 4.32)	1.74 (1.38, 2.30)	Lognormal	4.53 (4.03, 5.11)	2.68 (2.22, 3.34)	0.55 (0.40, 0.66)
	Community	76	Clayton	Gamma	4.45 (3.87, 5.12)	2.36 (1.85, 3.23)	Lognormal	4.99 (4.40, 5.68)	2.93 (2.42, 3.69)	0.48 (0.30, 0.64)
Basis for selection	Cluster	138	Gumbel	Gamma	4.09 (3.63, 4.61)	2.18 (1.73, 2.81)	Lognormal	4.55 (4.13, 5.03)	2.50 (2.13, 3.00)	0.46 (0.29, 0.59)
	Contact pattern	54	Clayton	Weibull	4.63 (4.03, 5.30)	2.31 (1.82, 3.10)	Lognormal	5.13 (4.45, 5.91)	3.07 (2.46, 3.95)	0.59 (0.42, 0.68)
	Domestic travel	43	Clayton	Gamma	4.84 (4.13, 5.65)	2.64 (2.00, 3.58)	Lognormal	5.00 (4.21, 5.91)	3.02 (2.40, 3.89)	0.47 (0.23, 0.64)
	Import	22	Gaussian	Weibull	4.83 (3.91, 5.79)	2.33 (1.61, 3.48)	Lognormal	5.66 (4.60, 6.78)	3.49 (2.72, 4.55)	0.65 (0.37, 0.69)
Asymptomatic infectors		29	-	Lognormal	4.62 (3.85, 5.45)	2.27 (1.70, 3.22)	-	-	-	-

CrI: credible interval. Wave 1 began 16 January 2020. Wave 2 is assumed to have begun 1 June 2020, and wave 3 is assumed to have begun 1 October 2020.

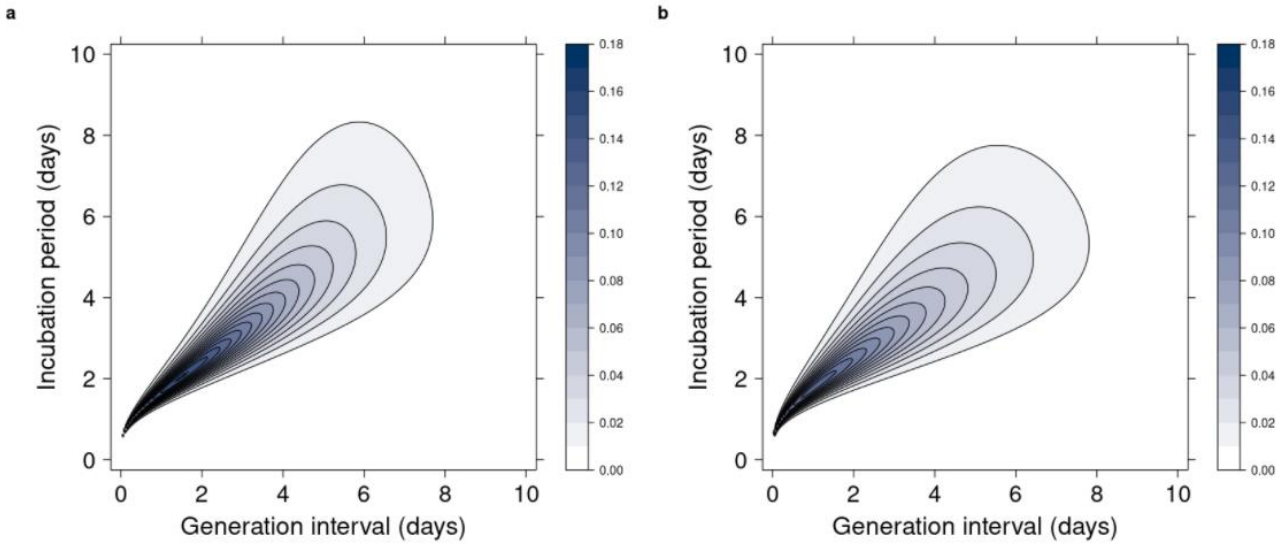

Figure 4. **Joint distribution of the generation interval and incubation period.** Contour plots of the fitted distributions. For both a, the dataset of 49 transmission pairs with single dates of possible exposure, and b, the dataset of 257 transmission pairs that also includes pairs with more coarsely reported possible dates of exposure and contact, a Clayton copula with a Weibull marginal for the generation interval and lognormal marginal for the incubation period distribution was selected.

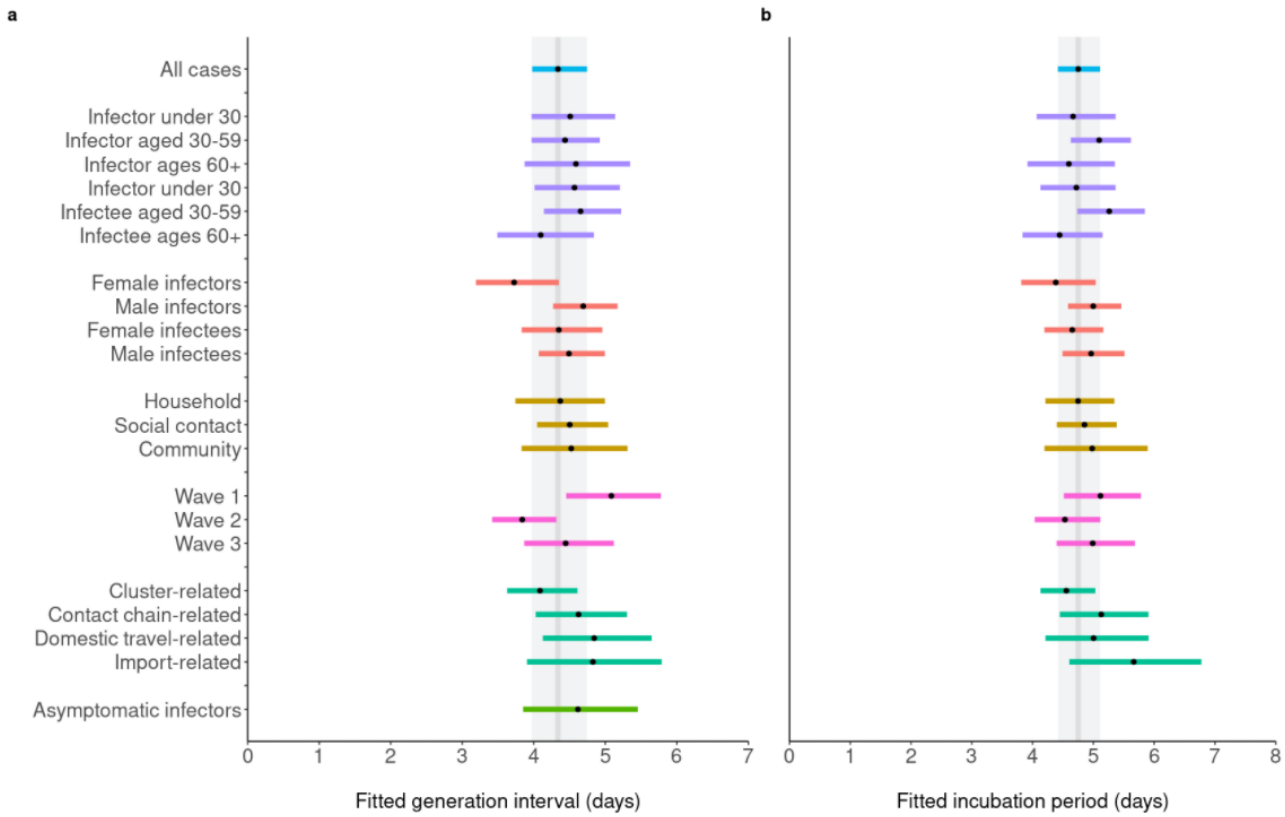

Figure 5. Joint estimates of the generation interval and incubation period by subgroup for COVID-19 transmission pairs from Japan. The joint distribution using best-fit Gaussian, Gumbel, or Clayton copula combined with gamma, lognormal, or Weibull distributions for the a, generation interval and b, incubation period are presented for the dataset of 257 transmission pairs that also includes pairs with more coarsely reported possible dates of exposure and contact. The points are point estimates for the means of each subgroup, while the colored bars indicate the 95% credible intervals. The grey bars show the overall point estimate and 95% CrI for all cases in the background. The estimate for asymptomatic infectors was fitted to the generation interval alone, as infector incubation period could not be estimated.

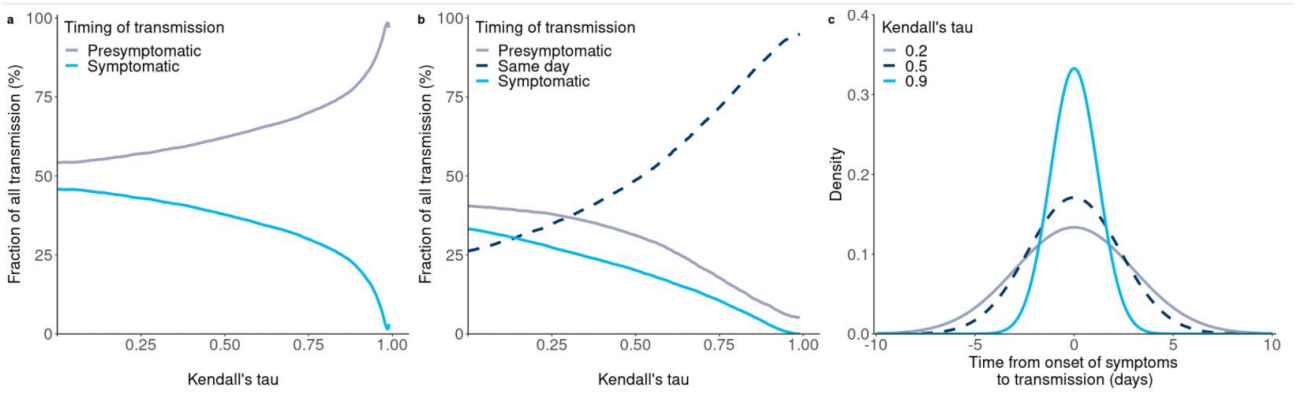

Figure 6. Increased correlation leads to a predominance of presymptomatic transmission. From estimates made by simulating the generation interval (GI) and incubation period (IP) for 10000 pairs using the fitted Clayton copula with Weibull (GI) and lognormal (IP) marginals: a, the proportion of transmission that was symptomatic or presymptomatic for various values of Kendall's tau; b, the proportion of transmission that was symptomatic, presymptomatic, or occurred on the same day (GI-IP $\in [-\mathbf{1}, \mathbf{1}]$) for various values of Kendall's tau; c, the time from onset of symptoms to transmission, defined as GI-IP, fitted with a normal distribution.

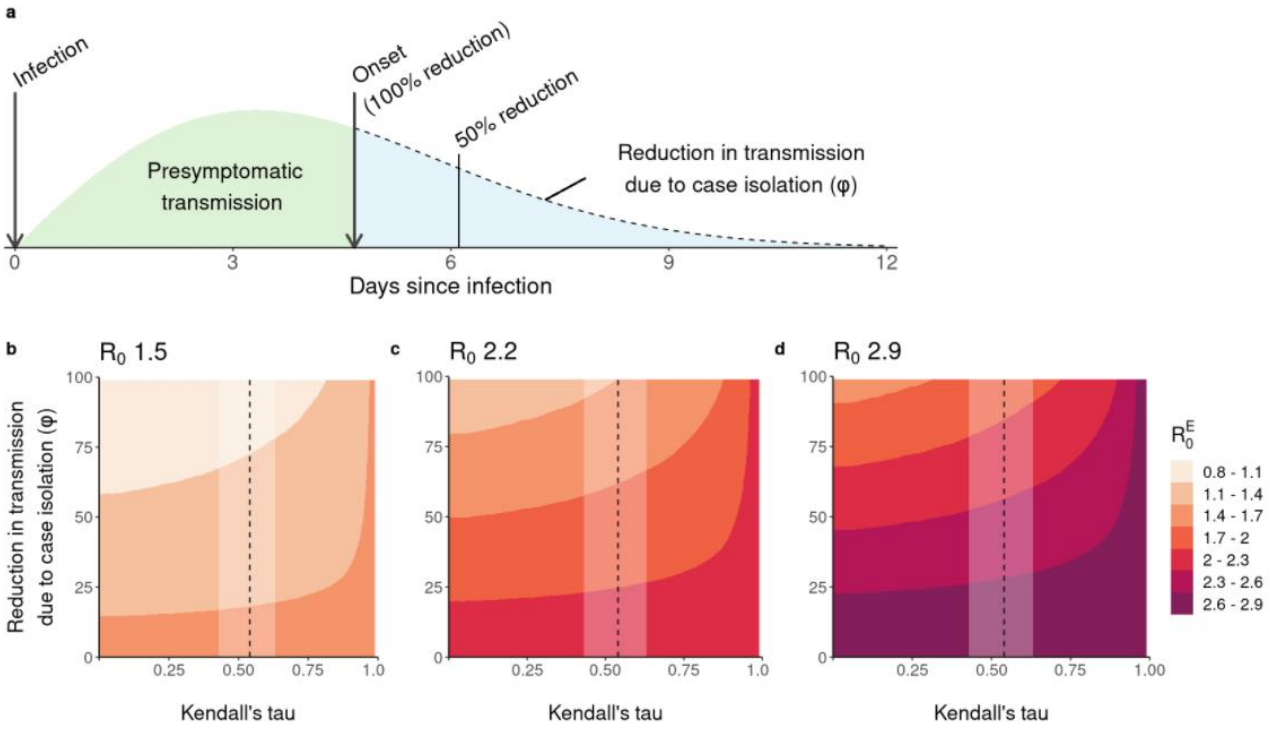

Figure 7. Effect of stronger correlation between the generation interval and incubation period on effectiveness of isolation. The top figure, a, shows the fitted generation interval probability distribution function (4.22 days). The arrow dividing the green and blue sections indicates onset at 4.67 days (the mean incubation period). If case isolation occurs at onset this is equivalent to a 100% reduction in possible transmission for symptomatic cases. Figures b, c, and d, show the effective basic reproduction number (R_0^E) as a function of this reduction in transmission as well as the level of correlation between the generation interval and incubation period. The dashed line is the point estimate of Kendall's tau obtained in this study, while the shaded white rectangle shows its 95% credible interval. We assume baseline R_0 of b, 1.5, c, 2.2, and d, 2.9. As the generation interval and incubation period approach independence (Kendall's tau→ 1) case isolation will become ineffective—shown by the unchanging effective R_0 (R_0^E)—as transmission will either be presymptomatic or occur nearly at the same time as symptom onset.

Correlation, presymptomatic transmission, and control measures

We found that 63.2% of pairs experienced presymptomatic transmission, defined as the generation interval being shorter than the incubation period, by simulating 10,000 transmission pairs from our fitted estimates of the generation interval, incubation period, and Kendall's tau (Table 2). Using this fit, we varied Kendall's tau for the same estimates of the generation interval and incubation period and show that as Kendall's tau approached zero (independence) the proportion of presymptomatic transmission reached a lower boundary of 54.2%. Conversely, as Kendall's tau approached 1 (complete dependence), the proportion of presymptomatic transmission increased to nearly 100% (Figure 3a), and the difference between symptom onset and transmission became so small that they mostly occurred on the same day, with only a small portion of presymptomatic transmission occurring outside of the 24 hours before or after symptom onset, and no symptomatic transmission occurring.

The average difference between the generation interval and incubation period across all data subsets and subgroups was 0.4 days, indicating a mean time from onset of symptoms to transmission of -0.4 days. The probability density function of the time from onset of symptoms to transmission fitted with a normal distribution based on simulated data with Kendall's tau varied between 0.2, 0.5, and 0.9, is shown in Figure 3c. The mean was centered at -0.4 days, and lower correlation resulted in a larger standard deviation.

Using the same simulated dataset, Kendall's tau was varied against the fraction of transmission reduced by case isolation. Increasing Kendall's tau indicated greater difficulty in controlling transmission via isolation alone given the same level of reduction in transmission due to rapid isolation φ (Figure 4). Given a basic reproduction number of R_0 of 2.2 or 2.9 (Riou and Althaus, 2020), the effective R_0 (denoted R_0^E) failed to reduce below the epidemic threshold of 1 (Figure 4b and c). Moreover, as Kendall's tau approaches 1 and the proportion of presymptomatic transmission (with a mean time from onset of symptoms to transmission of approximately zero) approaches 100%, control through isolation alone becomes impossible, and there is no difference between R_0 and R_0^E. When case isolation does not occurs promptly following onset (lower φ) the

dependence between the generation interval and incubation period has less impact on the reduction of R_0^E.

Discussion

The generation interval underpins many infectious disease models (Cori et al., 2013; Gostic et al., 2020), and here we provide insight into the generation interval of COVID-19 over time and across different characteristics of transmission pairs, providing one of the most comprehensive characterizations of the generation interval of COVID-19 to date. In addition, we quantitatively measured the relationship between the generation interval and incubation period of COVID-19, the lack of which was identified as a limiting factor in previous studies (Kremer et al., 2020; Lehtinen et al., 2021; Tindale et al., 2020). From transmission pairs identified using publicly available data reported in Japan during 2020 we found positive correlation between the generation interval and incubation period with a Kendall's tau ranging from 0.4–0.7. The mean generation interval was consistently shorter than the mean incubation period when jointly estimated, with the former ranging between 3.7–5.1 days and the latter between 4.4–5.7 days, indicating consistent presymptomatic transmission.

The means of the jointly estimated generation interval and incubation period are in line with those reported elsewhere, as the mean generation interval reported here—4.3 days—falls between the range of 2.8–7.5 days previously reported; see Supplementary Table 1 and Figure S8 (Ferretti et al., 2020; Ganyani et al., 2020; Li et al., 2021; Tindale et al., 2020). The positive correlation between the generation interval and incubation period indicates that for symptomatic cases, onset is tied to infectiousness. This finding supports evidence shown in virological studies (He et al., 2020b). The estimate of the generation interval for asymptomatic infectors was longer than that estimated for symptomatic infectors, at 4.7 days, indicating that infectiousness in asymptomatic cases may be more persistent than in symptomatic infections—potentially leading to an underestimation of R_0 using estimates from symptomatic pairs (Park et al., 2020c). Although evidence has indicated that asymptomatic COVID-19 cases are less infectious than symptomatic cases (Nakajo and Nishiura, 2021), asymptomatic cases nonetheless play a role in epidemic dynamics (Ferretti et al., 2020).

The proportion of presymptomatic transmission among symptomatic cases estimated in this study, 63.2%, is higher than estimates reported in some early publications (Ferretti et al., 2020; He et al., 2020b), but similar to others.(Ganyani et al., 2020; Hart et al., 2021; Tindale et al., 2020) Among the 63.2% of simulated pairs with presymptomatic transmission, 33.5% had transmission occur within one day of infector onset, and 49.1% within two days of infector onset. Combining estimates of asymptomatic transmission in the range of 18–30% (Mizumoto et al., 2020; Nishiura et al., 2020b) with the estimate of presymptomatic transmission shared here, the proportion of nonsymptomatic transmission could feasibly reach 90%. Thus, interventions such as physical distancing that do not depend on detection of potential infectors while they are not showing symptoms, enhanced surveillance to detect nonsymptomatic cases, and contact tracing to identify exposed individuals while their infected contacts are not symptomatic are crucial for COVID-19 control efforts (Ferretti et al., 2020; Fraser et al., 2004; Lovell-Read et al., 2021).

Data collection for this study focused on high certainty of directionality of transmission based on publicly shared epidemiological data. In contrast to most other countries, Japan applied backward contact tracing methods from the beginning of the pandemic in an effort to prevent large clusters of cases (Oshitani and The Expert Members of the National COVID-19 Cluster Taskforce at the Ministry of Health Labour and Welfare of Japan, 2020), making it an apt setting for obtaining transmission pair data. This is because links between cases—and particularly those related to clusters—were more likely to have been detected compared to countries where backward contact tracing was not conducted. However, the timing of COVID-19 testing plays an important role in case ascertainment (Long et al., 2021), and infected persons who were epidemiologically linked to COVID-19 cases but did not become symptomatic after initially testing negative for SARS-CoV-2 may been missed as cases. In addition, Japan did not promote widespread community viral testing, and this perhaps limited the number of unlinked cases that may have otherwise been detected and retrospectively linked to others during the epidemic. As well, public health jurisdiction reporting practices changed over time, with details generally becoming sparser once daily incidence became high enough to wear contact tracing capacity thin, and also towards the end of the year as concerns grew around infection-related stigmatization (Yoshioka and Maeda, 2020).

The shorter mean generation interval during the second and third waves of pandemic in Japan compared to the first wave may in part reflect the increase in prevalence of infection, as increased competition between infectious individuals to find susceptible contacts can lead to contraction of the generation interval (Ali et al., 2020; Kenah et al., 2008; Nishiura, 2010). Shorter generation intervals were also noted in the United Kingdom during September–November 2020, when there was a rise in the number of new cases (Hart et al., 2021). However, our results indicate that increases in incidence do not perforce lead to contraction of the generation interval. The larger value obtained for the third wave in Japan, which had a higher peak than the second wave (Figure S9), may reflect an aspect of increased heterogeneity in transmission not captured in our analysis, such as community awareness of the dangers of COVID-19 and local case count, as well as government-implemented public health and social measures.

During the first epidemic wave, state of emergency declarations nationwide. However, during the second wave and the 2020 half of the third wave, no such preventative measures were instituted. Conversely, campaigns intending to restart the Japanese economy following the difficulties caused by the first wave of COVID-19 were developed and implemented. In particular, the GoTo Travel campaign, which offered discounts on travel inside Japan, was a fixture of the second and third waves (Figure S9). The campaign began just before the peak of the second wave and was associated with an increase in COVID-19 cases reporting inter-prefecture travel (Anzai and Nishiura, 2021). Of our pairs identified for the second wave, only 20.0% of infectors were reported before the start of the GoTo travel campaign. Although we did not find that our pairs where the infector had domestic travel experienced longer generation intervals (**Figure 5**), travel can left-censor the time following infection when an infector who travelled had contact with an infectee who did not travel with them.

Most identified transmission pairs had contact in the household or in settings related to social behavior, such as eating at restaurants, visiting nightlife, singing karaoke, attending sports events, listening to live music, visiting gyms, or meeting with friends, relatives, acquaintances, etc. (Table 6). These types of social contact settings have also been associated with SARS-CoV-2 transmission in other countries (Leclerc et al., 2020). In Japan, settings for social contact first

are the first to be requested to be restricted by prefectural and local governments when control measures or a state of emergency was deemed necessary to reduce case incidence (Cabinet Secretariat of Japan, 2021), though such emergency measures were not implemented during the second or third wave portions of 2020. Similar interventions focused on limiting social contact were implemented in other parts of the world (Imai et al., 2020a), though other countries had a greater focus on reducing formalized community contact, such as by moving schools and workplaces online (Brauner et al., 2021; Liu et al., 2020).

As SARS-CoV-2 variants begin to dominate transmission in many countries (Davies et al., 2021; Ito et al., 2021; Ministry of Health Labour and Welfare (MHLW), 2021), it remains to be seen whether the mean and variance of the generation interval for the new variants of concern (VOC) will be similar to estimates presented in this, or previous studies. It has been suggested that the Alpha (B.1.1.7) variant could have either shorter or longer generation intervals (Davies et al., 2021; Gaymard et al., 2021; Kissler et al., 2021), and may therefore be responsive to interventions that were not successful with wild-type SARS-CoV-2, or vice-versa (Park et al., 2020d). If SARS-CoV-2 variant generation intervals are longer and the viral load is higher, this may also lengthen the tail of the generation interval, increasing the possibility for onward transmission by undetected cases without interventions that target nonsymptomatic transmission (Kissler et al., 2021). Whether the correlation between the generation interval and incubation period would be weaker or stronger than has been presented here remains to be seen, however this study provides a basis for consideration of such correlation moving forward.

Supplementary materials

Previous estimates of the generation interval of COVID-19

Some early estimates of the generation interval of COVID-19 were published during the first half of 2020 (Ferretti et al., 2020; Ganyani et al., 2020; Tindale et al., 2020). However, they were not estimated directly from data on exposure dates but derived from estimates of the serial interval and incubation period, as indicated in Figure S8 and Table S2 below. We found only one other study (Li et al.) that estimated the generation interval directly from reported dates of exposure, although this study assumed infection occurred at the middle of a given exposure period (Li et al., 2021), which may have been intended to address the censored observational data. None of the studies specifically aimed to estimate the generation interval for pairs with asymptomatic infectors.

- Ferretti et al. used transmission pairs selected based on high confidence of direct transmission inferred from publicly available sources.

- Ganyani et al. used datasets of all reported cases between 21 January to 26 February 2020 (Singapore) and 14 January to 27 February 2020 (Tianjin). For cases linked to clusters, they imputed links between cases to ascertain pairs.

- Tindale et al. used the same datasets as Ganyani et al. Linkages, when not explicitly available from the data, were established using the methods described in te Beest et al. (Te Beest et al., 2013).

- Bushman et al. combined transmission pair data from four published studies (Du et al., 2020; He et al., 2020b; Xu et al., 2020; Zhang et al., 2020a).

- Li et al. limited their analysis to transmission pairs where the infector had travel history to Hubei Province, China (where the original epicenter of COVID-19, Wuhan, is located). Of note, they did not apply doubly-interval censoring, as has been done

elsewhere (Backer et al., 2020; Linton et al., 2020; Reich et al., 2009), but assumed infection occurred at the exact middle of a given exposure period.

- Hart et al. estimated the generation interval from cases reported in households in the United Kingdom. Their methods considered asymptomatic infectors, and they considered a model where infectiousness and symptom onset were independent, as well as a model (their “mechanistic” model) where infectors who developed symptoms progressed through different stages of infection.

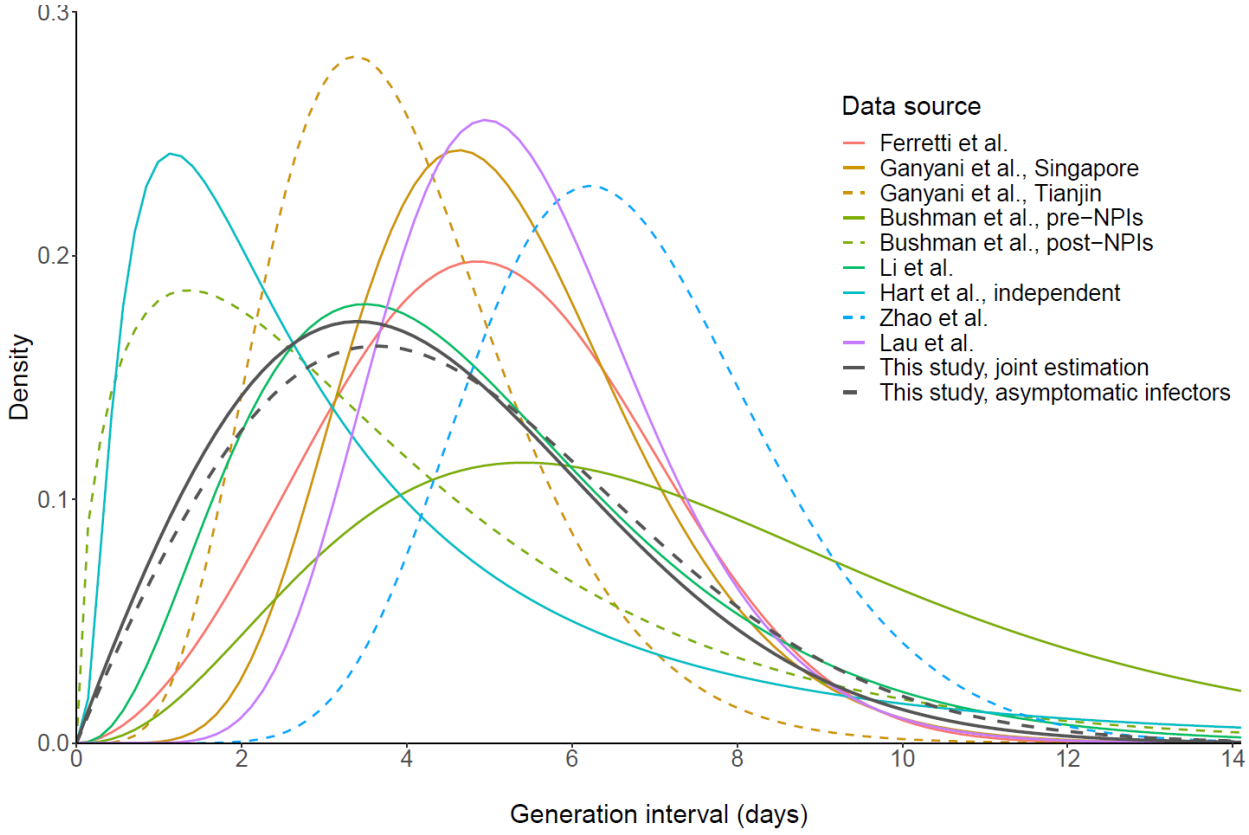

Figure S8. Generation intervals reported across COVID-19 studies. The studies by Ganyani et al., Bushman et al., Lau et al., Li et al., and Zhao et al. are plotted using gamma distributions. The result from Hart et al. was plotted using a lognormal distribution. Results from this study and the study by Ferretti et al. plotted using Weibull distributions.

Table S2. Estimates of the generation interval of COVID-19 from previously published studies.

Estimation method	Mean (95% CI)	SD (95% CI)	Distri-bution	Geographic scope	Pairs	Reference
From SI and IP†	5.04 (4.19, 6.31)	1.93 (1.52, 2.47)	Weibull	Worldwide	40	(Ferretti et al., 2020)
From IP intermediates	3.71 (2.36, 4.91)	-	Gamma	Singapore	56*	(Tindale et al., 2020)
From IP intermediates	2.82 (1.82, 3.52)	-	Gamma	Tianjin, China	72*	Tindale
From SI and IP‡	5.20 (3.78, 6.78)	1.72 (0.91, 3.93)	Gamma	Singapore	-	(Ganyani et al., 2020)
From SI and IP‡	3.95 (3.01, 4.91)	1.51 (0.74, 2.97)	Gamma	Tianjin, China	-	Ganyani
From SI and IP†	7.50 (6.81, 8.20)	3.95 (3.32, 4.74)	Gamma	China, pre-NPI	873	(Bushman et al., 2021)
From SI and IP†	3.90 (3.59, 4.24)	3.15 (2.78, 3.53)	Gamma	China, post-NPI	873	Bushman
Directly	4.81 (4.13, 5.58)	2.52 (1.93, 3.32)	Gamma	China	67	(Li et al., 2021)
Independent§	4.2 (3.3, 5.3)	4.9 (3.0, 8.3)	Gamma	United Kingdom	172¶	(Hart et al., 2021)
Mechanistic§	6.0 (5.2, 7.0)	4.9 (4.0, 6.3)	Gamma	United Kingdom	172¶	Hart

CI: confidence/credible interval; COVID-19: coronavirus disease 2019; IP: incubation period; NPI: non-pharmaceutical intervention; SD: standard deviation; SI: serial interval. *Manually counted from Fig. 6. †IP distribution based on (Lauer et al., 2020). ‡IP distribution based on (Zhang et al., 2020a). §IP distribution based on (McAloon et al., 2020). ¶172 households with 603 cases.

Among these studies, only Bushman et al. and Hart et al. considered dependence between the generation time and incubation period. The mechanistic model of Hart et al. considers that symptoms and transmission are not independent, but not directly quantify possible dependence. Rather, the model conditions infectiousness on the duration of the incubation period. Bushman et al., although they considered "incubation-dependent" models, found that the best fits for their data were "incubation-independent" models, suggesting low- or no correlation between the generation interval and incubation period. Their method for defining "incubation-dependent" models was to vary the rate parameter of the gamma distribution by dividing it by the length of the incubation period, and it is possible this method—as it did not include correlation as a parameter—could not capture existing correlation. As well, it is possible that the method they used to calculate the generation interval from the serial interval—which assumes independence from the incubation period (Bushman et al., 2021; Ferretti et al., 2020; Ganyani et al., 2020; Lehtinen et al., 2021)—alters the underlying correlation structure.

Table S3 lists previously published studies where correlation between transmission intervals (generation or serial interval) and the incubation period were assessed. Dependence between the two parameters has rarely been estimated.

Klinkenberg and Nishiura (2011) estimated the correlation between the generation interval and incubation period of measles in Rhode Island prior to development of the measles vaccine (Table S3). They found very different results depending on whether they were using bigamma or bilognormal marginals. The better fit to the bilognormal marginals may reflect the better fit to the lognormal distribution commonly seen in incubation period data for respiratory diseases (Lessler et al., 2015), including COVID-19 (Backer et al., 2020; Lauer et al., 2020; Linton et al., 2020). Tindale et al. assessed covariance between the serial interval and incubation period of COVID-19 using empirical data from their selected transmission pairs. However, the pairs ignored negative serial intervals.

Table S3. Estimates of correlation between transmission intervals and incubation periods from other studies

Disease	Transmission interval	Kendall's tau (CrI or p-value)	Geographic scope	Time period	Method	Reference
Measles	Generation	0.27 (-0.34, 0.76)*	Rhode Island	1917–1923	Copula, gamma	Klinkenberg (Klinkenberg and Nishiura, 2011)
Measles	Generation	0.63 (0.14, 0.88)*	Rhode Island	1929–1934	Copula, gamma	Klinkenberg
Measles	Generation	0.72 (0.52, 0.84)*	Rhode Island	1917–1923	Copula, lognormal	Klinkenberg
Measles	Generation	0.84 (0.67, 0.95)*	Rhode Island	1929–1934	Copula, lognormal	Klinkenberg
COVID-19	Serial†	0.13 (p=0.2)	Singapore	2020	Case pairs	Tindale (Tindale et al., 2020)
COVID-19	Serial†	0.19* (-)	Tianjin, China	2020	Case pairs	Tindale

CrI: credible interval; COVID-19: coronavirus disease 2019. *Original results were presented as correlation coefficient ρ, which is defined in terms of Kendall's tau as $\tau_K = \frac{2}{\pi}\arcsin(\rho)$. †Estimates of the serial interval included only positive values.

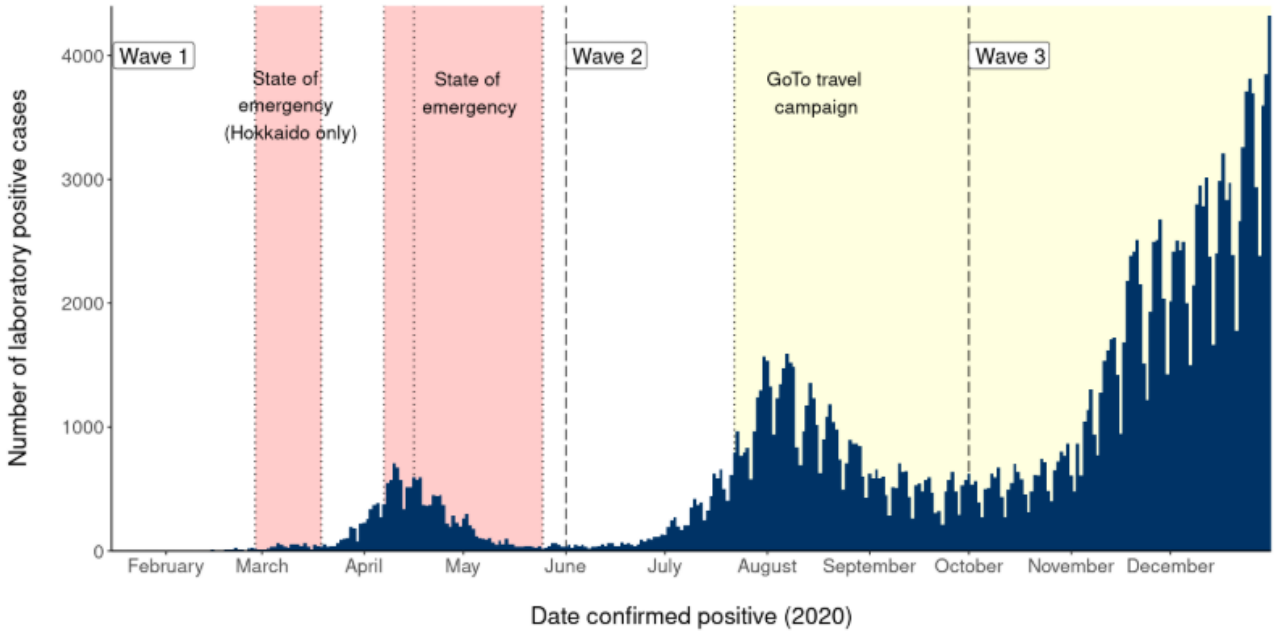

Figure S9. Epidemic curve of COVID-19 cases in Japan by date of laboratory confirmation. Here, the second epidemic wave is shown as beginning 1 June (first dashed line), while the third epidemic wave starts 1 October (second dashed line). Hokkaido declared a local state of emergency 28 February–19 March (first two vertical dotted lines). A national state of emergency was declared for key urban prefectures (Tokyo, Saitama, Chiba, Kanagawa, Osaka, Hyogo, and Fukuoka) on 7 April (third dotted line), with the state of emergency extending nationwide on 16 April (fourth dotted line). The end of the state of emergency varied between prefectures, with most ending 14 May while some continued until 25 May (fifth dotted line). The GoTo travel campaign, offering large discounts on travel inside Japan with the intention of restarting the Japanese economy following the damage caused by COVID-19 related public health and social measures, began on 22 July (sixth dotted line) for all prefectures except Tokyo, which was added to the campaign on 1 October (start date of third wave, second dashed line). The campaign continued through the end of 2020 and into 2021.

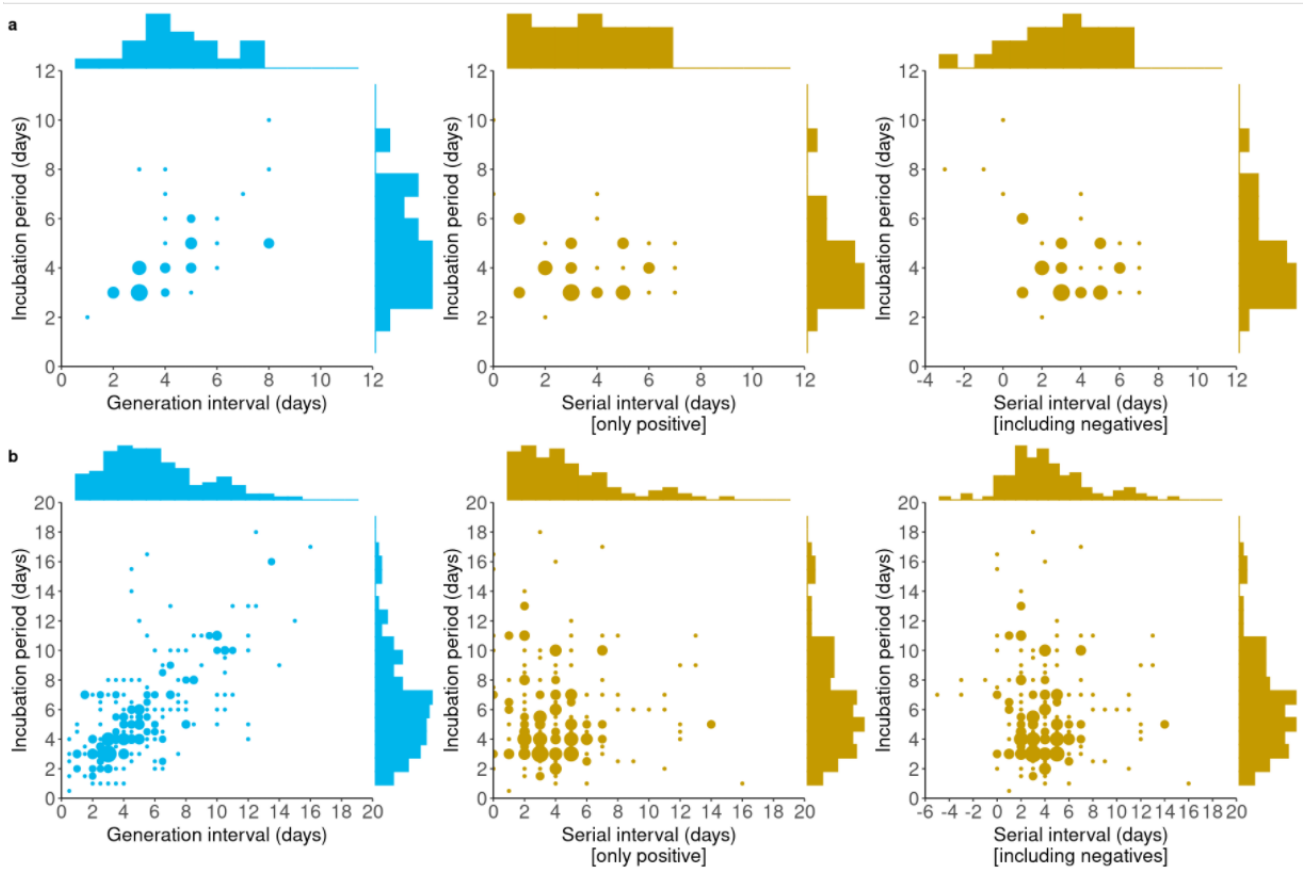

Figure S10. Correlation of empirical generation and serial intervals with the incubation period of COVID-19 cases in Japan. Scatterplots showing the generation and serial intervals plotted against the incubation period for transmission pairs with single-date (n=49) and coarsely observed (n=257) exposures. For **a**, data where single dates of exposure for the infector and contact between infector and infectee were reported, Kendall's tau was 0.58 ($p<0.001$), -0.13 ($p=0.32$), and -0.23 ($p=0.07$) for the generation interval, serial interval with only positives, and serial interval including negatives, respectively. For **b**, data with coarse dates of exposure and contact, the plotted value represents the midpoint of the possible exposure/contact period, and Kendall's tau was 0.53 ($p<0.001$), -0.02 ($p=0.66$), and -0.04 ($p=0.39$).

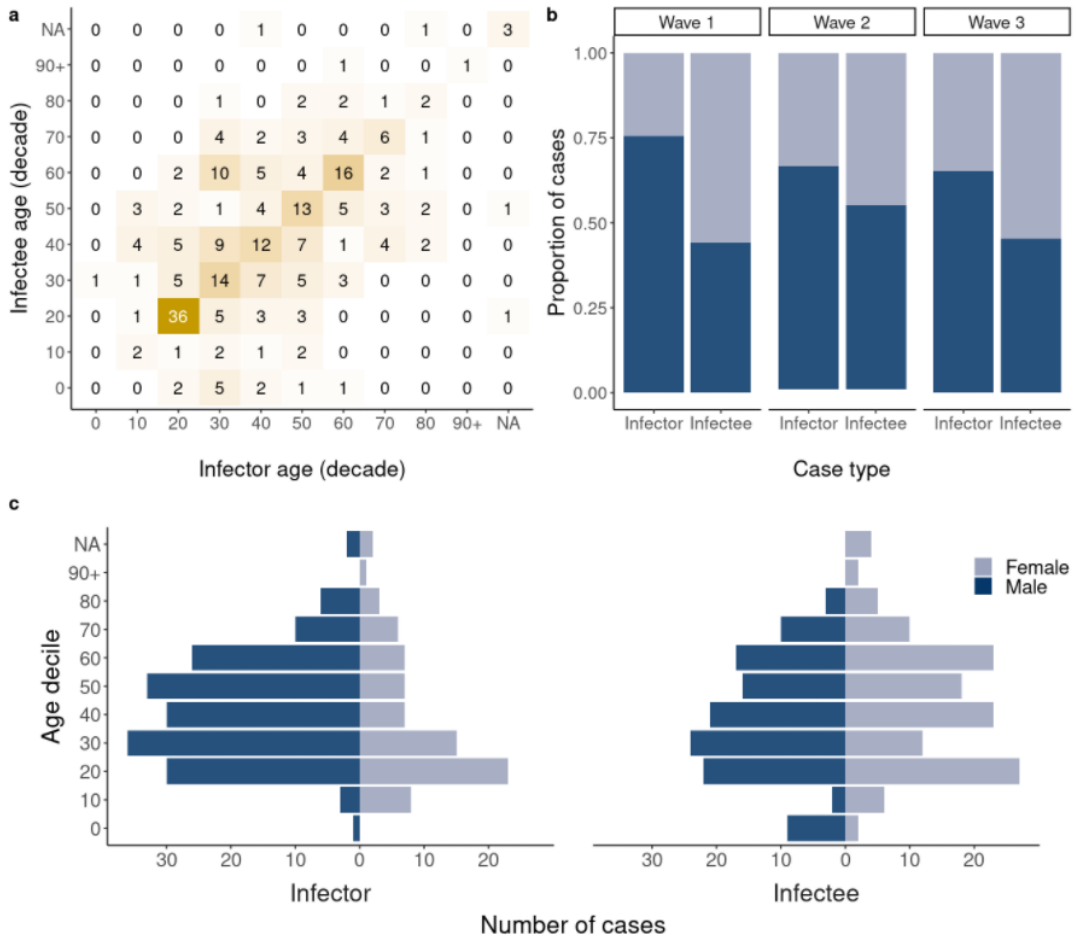

Figure S11. Age and sex distribution of transmission pairs. a, Matrix of infector and infectee age groups for 257 transmission pairs. b, Proportion of cases by sex, wave, and infector-infectee status. **c**, Age group and sex distribution of infectors and infectees.

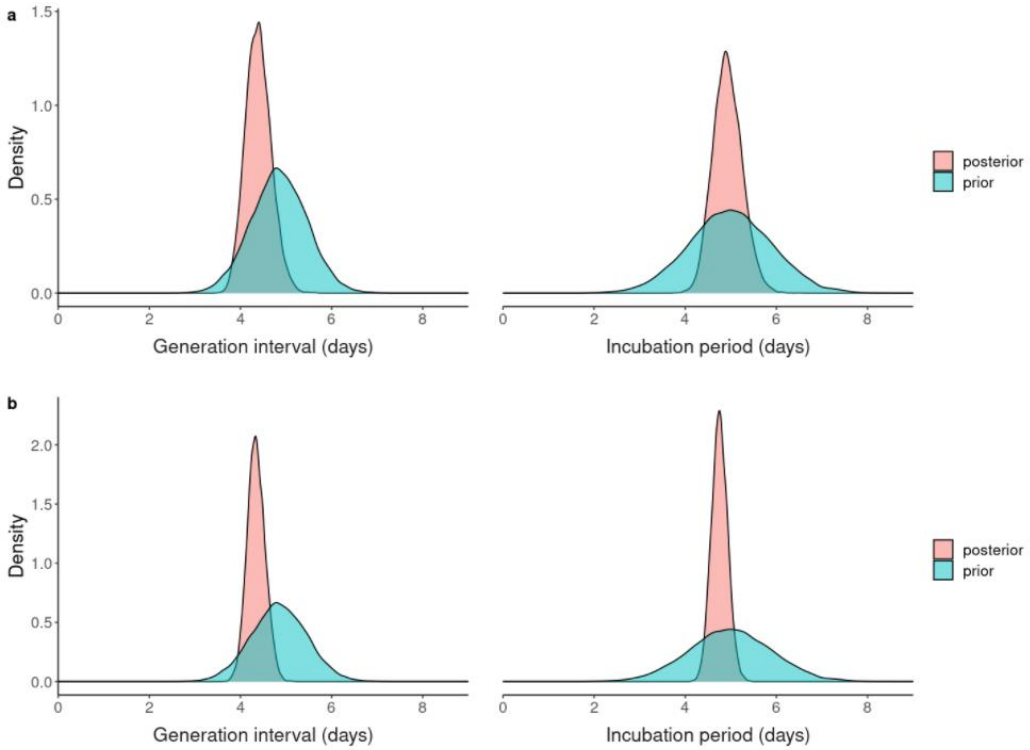

Figure S12. Prior and posterior densities of the mean. a, densities for the dataset where single dates of exposure for the infector and contact between infector and infectee were reported (n=49). **b**, densities for the dataset with coarsely reported dates of exposure and contact (n=257). The priors are the same for **a** and **b**; for the generation interval the prior was obtained from the values of the serial interval reported by Nishiura et al., while the prior for the incubation period was obtained from the values of the incubation period reported by Linton et al.—both studies used cases reported worldwide during the early stages of the pandemic (early 2020).

Case definition and determination of transmission pairs

Nearly all COVID-19 cases in Japan had a positive viral test for SARS-CoV-2, with a few exceptions made for cases with positive antibody tests based on clinical judgement. These viral tests were either nucleic acid amplification tests (NAATs) or antigen tests. Not all public health jurisdictions shared the type of test in their case reports, but typically the NAATs were either reverse transcription polymerase chain reaction (RT-PCR) or loop-mediated isothermal amplification (LAMP) tests.

Presymptomatic and asymptomatic transmission are possible for SARS-CoV-2 infections (He et al., 2020b; Tindale et al., 2020). Thus, directionality of transmission for epidemiologically linked cases that is determined solely based on dates of onset among linked cases is likely to include some misclassification of infector and infectee status and may ignore possible intermediate infectors. We collected information on COVID-19 cases reported in Japan and looked for epidemiological information that would provide insight into directionality of transmission between linked cases. Criteria for ascertainment of directionality of transmission generally fell into one of four categories:

1. Import: If a potential infector had onset during or following international travel to a country with COVID-19 cases they were considered an imported case.
2. Cluster: A cluster was defined as five or more cases. If a potential infector was linked directly or via a chain of infections to a cluster (common exposure) the link was classified as cluster-based.
3. Domestic travel: If onset of an infector occurred during or within 10 days after travel to another prefecture and there were no more obvious possible exposures these cases were labelled as having domestic travel as their possible exposure, and the dates of travel form the left- and right-hand bounds of their exposure.
4. Contact pattern: the contact pattern between cases, typically supplemented by some reported dates of contact or exposure, provided insight into directionality of infection for linked infector-infectee pairs.

We supply a variable “link_basis” in the dataset of transmission pairs as a means of assessing whether there was any difference between estimates given the differing rationales for ascertaining directionality of transmission for pairs for each of the above categories.

Cleaning dates of exposure and contact

Reported dates on exposure, contact, and symptom onset were cleaned to obtain EL and ER (left- and right-hand exposure times for the infector), CL and CR (left- and right-hand times of contact between the infector and infectee), as well as S1 and S2 (symptom onset times of the infector and infectee). All times are reported as number of days relative to a “time zero” of 1 January 2020. An abridged data dictionary for calculated values is provided in Table S4.

For most case reports that included temporal information on potential exposures or contact between cases, exact dates or date ranges were explicitly stated. However, some of the data may include inferences made from statements related to parts of the month, with data entry performed as follows:

- Beginning of the month (月初) = until the seventh day of the month. More common may be just the first day of the month, or until the third day.
- End of the month (月末) = from the 25th day of the month. An informal survey found that approximately one-third of respondents felt that the phrase “end of the month” could include dates as early as the 25th (see https://mainichi-kotoba.jp/enq-255).
- Beginning, middle, and end of the month (上旬, 中旬, 下旬) = days 1–10, 11–20, 21–last day of a given month, respectively.

In cases where left- or right-hand bounds for exposure and contact were missing from the data, we used other epidemiological information to substitute for these bounds when plausible to do so. The assumptions were as follows:

1) Data cleaning applied to ER/CR:
   a) For some cases, recorded right-hand exposure (e.g., travel, contact with a case) for infector and/or infectee may exceed symptom onset (S1 or S2). As we are only interested in exposure and contact dates as they related to the possible infection time for infector and infectee, we set ER and CR equal to S1 and S2.

b) Similarly, for some infectors, their recorded right-hand exposure may exceed their time of contact with the infectee. In these scenarios we presume the infector is infected by the time of final contact the infectee, so we censor ER to CR.
c) Missing ER were sequentially assigned to the minimum of either:
    i) S1,
    ii) CR,
    iii) date of infector laboratory confirmation,
    iv) date of infector isolation,
    v) or 14 days* after EL.
d) Missing CR were sequentially assigned to the minimum of either:
    i) S2,
    ii) date of infectee laboratory confirmation,
    iii) or 14 days* after CL.

2) Data cleaning applied to EL/CL:
    a) Missing EL were assigned to 14 days* before ER.
    b) In some cases, infectees may have been reported to have contact with the infector prior to the infector's left-hand exposure (EL). For example, the infector and infectee may have met several times. It is possible that the left-hand exposure for the infectee (CL) was assigned to a date that was earlier than EL. However, as pairs included in the dataset were selected for having a relatively high likelihood of the directionality of infection being true, in cases where CL < EL, we set CL = EL.
    c) If the infector exposure type was labelled "International travel" and the infectee was missing CL, then CL was set to ER.
    d) If contact type between infector and infectee is "household," link basis was "cluster" or "domestic contact", infectee was missing CL, and EL was not missing, CL was set to EL.
    e) Missing CL were assigned to EL, as an infector cannot be infectious (and therefore CL cannot be a valid date of contact for transmission to occur from infector to infectee) unless the infector's own transmission has occurred.

*14 days was selected as this ~>95% of the incubation period (Backer et al., 2020; Lauer et al., 2020; Linton et al., 2020). Although the above assumptions made it possible to obtain EL and

ER even in a left- or right-hand bound of exposure was not reported, we only included cases where a left- or right-hand bound of exposure was explicitly reported.

Table S4. Calculated variables included in the dataset.

Variable	Description	Class	Values	Details
EL	Left-hand bound of infector exposure	Numeric	Days	
ER	Right-hand bound of infector exposure	Numeric	Days	
CL	Left-hand bound of infectee contact with infector	Numeric	Days	
CR	Right-hand bound of infectee contact with infector	Numeric	Days	
S1	Symptom onset day of infector in days	Numeric	Days	
S2	Symptom onset day of infectee in days	Numeric	Days	
link_basis	What the basis was for determining directionality of transmission between linked pairs	Factor	Cluster	Infector was linked to a cluster or part of a transmission chain linked back to a cluster
			Contact pattern	Timings of contact and onset between cases provide plausible evidence for directionality of transmission
			Domestic travel	Infector travelled domestically to a location with ongoing transmission and timing of travel is plausibly related to onset of disease
			International travel	Infector travelled internationally part of a transmission chain linked back to international travel, and said travel is believed to have been the source of infection
exposure_type	Transmission setting for infector exposure.	Factor	Household	A household member or family member (when household status was not specified)
			Social-contact based interaction	Venue for interaction is based on social interaction. Restaurants, nightlife, karaoke, sports events, live music, gyms, friends, relatives, acquaintances, etc., or type of contact is not specified.
contact_type	Transmission setting for contact between infector and infectee.		Core community interaction	Venue for exposure are schools, general workplaces, essential workplaces (care facilities, medical facilities, government services, etc.), or exposure is related to travel to another area and source of infection is unknown (community infection).

Bivariate joint distribution

We employed a Bayesian approach combining copulas (multivariate cumulative distribution functions) with doubly-interval censoring to obtain estimates of the generation interval and incubation period, as well as a measure of correlation those two parameters. For $i \in \{1, \dots, N\}$ transmission pairs, we obtain the following doubly-interval censored likelihoods for the generation interval and incubation period for data D:

$$L1\left(\Theta_j; \mathrm{D}\right) = \prod\nolimits_i \int_{E_{L.i}}^{E_{R.i}} \int_{C_{L.i}}^{C_{R.i}} j(e) f(\mathrm{c} - e) \mathrm{dcd}e\,, \tag{S7}$$

$$L2\left(\Theta_j; \mathrm{D}\right) = \prod\nolimits_i \int_{E_{L.i}}^{E_{R.i}} \int_{S_{L.i}}^{S_{R.i}} j(e) g(\mathrm{s} - e) \mathrm{dsd}e.$$

Here, e is the time of infection of the infector and $j(.)$ is the probability distribution function (PDF) of the time of infection of the infector following a uniform distribution across exposure time E_L to E_R. Similarly, c is the time of transmission of the pathogen from infector and infectee occurring between C_L and C_R, with $f(.)$ representing the PDF of the generation interval. Finally, s is the time of symptom onset of the infector occurring between S_L and S_R, with $g(.)$ representing the PDF of the incubation period.

Combining doubly-interval censoring with a copula function allowed us to obtain the bivariate joint distribution of the generation interval and incubation period. Copulas provide a correlation structure to sets of marginal distributions, allowing for the marginal distributions and dependence structure to be modeled separately. In accordance with Sklar's theorem (Sklar, 1959), the joint cumulative distribution function (CDF) can be decomposed into the copula and univariate marginal CDFs. As such, if $H(x, y)$ is a joint bivariate CDF with marginal CDFs $F(x)$ and $G(y)$, there exists a copula $C\colon [0,1]^2 \longrightarrow [0,1]$ such that:

$$H(x, y) = C\left(F(x), G(y)\right) \tag{S8}$$

for all $(x, y) \in [-\infty, \infty]^2$. From relations $F(x) = u$ and $G(y) = v$ we obtain $x = F^{-1}(u)$ and $y = G^{-1}(v)$. Substituting $F^{-1}(u)$ and $G^{-1}(v)$ into (S8) we obtain the copula:

$$C(u,v) = H(F^{-1}(u), G^{-1}(v)), \tag{S9}$$

for all $(u,v) \in [0,1]^2$. The marginal distribution functions are given by $F(x) = \int_{-\infty}^{x} f(u)du$ and $G(y) = \int_{-\infty}^{y} g(v)dv$. From the marginal distributions we have $U = F(X)$ and $V = G(Y)$ which are uniformly distributed in $[0,1]$. The joint distribution of (U,V) is the copula $C(u,v) = P(U \leq u, V \leq v)$. The density of the bivariate copula C is:

$$c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v} \tag{S10}$$

Setting x and $F(x)$ to be the data and CDF of the generation interval, and y and $G(y)$ to be the data and CDF of the generation interval, the joint distribution described in equation (S8) is therefore the joint distribution of the incubation period and generation interval. Given a copula parameter θ, the overall log-likelihood is expressed as the sum of the log-likelihood of each marginal distribution plus the log-likelihood of the copula:

$$\log L(\mu_X, \sigma_X, \mu_Y, \sigma_Y, \theta | x, y) = \log L(\mu_X, \sigma_X | x) + \log L(\mu_Y, \sigma_Y | y) + \log L(\theta | u, v). \tag{S11}$$

Copula selection

The Gaussian copula was utilized by Klinkenberg and Nishiura to assess correlation between the generation interval and incubation period for measles data (Klinkenberg and Nishiura, 2011), and therefore was of interest for inclusion in this study. However, the Gaussian copula does not account for tail dependence. To consider lower tail dependence we included the Clayton copula, while to consider upper tail dependence we included the Gumbel copula. The independence copula, which is the copula that results from a dependency structure in which each individual variable is independent of each other, was also considered. Table S5 introduces various properties of these four copulas.

Table S5. Bivariate copulas and their properties.

Copula	Copula parameter	Independence	Kendall's tau	Range of tau	Lower tail dependence	Upper tail dependence
Gaussian	$\theta \in (-1,1)$	$\theta = 0$	$\frac{2}{\pi}\sin^{-1}(\theta)$	[-1,1]	0	0
Gumbel	$\theta \in [1,\infty)$	$\theta = 1$	$1-\theta^{-1}$	[0,1]	0	$2-2^{\frac{1}{\theta}}$
Clayton	$\theta \in [-1,\infty)\backslash\{0\}$	$\theta \to 0$	$\frac{\theta}{\theta+2}$	[0,1]	$2^{-\frac{1}{\theta}}$	0
Independence	None	Always	0	0	0	0

Gaussian copula

The Gaussian copula is an elliptical copula. It is defined by

$$C(u,v) = \Phi_\theta\big(\Phi^{-1}(u), \Phi^{-1}(v)\big), \tag{S12}$$

where Φ is the standard normal distribution function and $\theta \in (-1,1)$ is the correlation between the components. The density of the bivariate Gaussian copula is given by

$$c(u,v;\theta) = \frac{1}{\sqrt{1-\theta^2}} exp\left\{-\frac{\theta^2(s^2+t^2)-2\theta st}{2(1-\theta^2)}\right\}, \tag{S13}$$

where $s = \Phi^{-1}(u)$ and $t = \Phi^{-1}(v)$. As such, the log-likelihood function for the Gaussian copula is

$$\log\big(L(\rho|u,v)\big) = -\frac{1}{2}\log(1-\theta^2) - \frac{\theta^2(s^2+t^2)-2\theta st}{2(1-\theta^2)}. \tag{S14}$$

For Gaussian copula, Kendall's tau is defined as $\frac{2}{\pi}\sin^{-1}(\theta)$. Independence is reached when $\theta = 0$.

Independence, Gumbel, and Clayton copulas

The Gumbel and Clayton copulas are single-parameter Archimedian copulas.(Czado, 2019) As a family, Archimedian copula are defined as

$$C(u,v) = \varphi^{[-1]}\big(\varphi(u) + \varphi(v)\big), \tag{S15}$$

where φ is the generator function of the copula C. $\varphi: [0,1] \to [0,\infty]$ is a continuous strictly decreasing convex function such that $\varphi(1) = 0$ and $\varphi^{[-1]}$ is the pseudo-inverse of φ, defined as $\varphi^{[-1]}: [0,\infty] \to [0,1]$ with

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t), & 0 \le t \le \varphi(0), \\ 0, & \varphi(0) \le t \le \infty. \end{cases} \tag{S16}$$

For the bivariate Gumbel copula, the generator is $\varphi(t) = (-log(t))^{\theta}$ and inverse generator $\varphi(t)^{-\theta} = \exp(-t^{1/\theta})$ for copula parameter $\theta \in [1, \infty)$, where t varies from 0 to 1 regardless of whether it is equal to u or v. The bivariate Gumbel copula is given as

$$C(u, v) = exp\left\{-\left[(-\ln(u))^{\theta} + (-\ln(v))^{\theta}\right]^{\frac{1}{\theta}}\right\}, \qquad \theta \in [1, \infty), \tag{S17}$$

where $\theta \to \infty$ indicates full dependence, while $\theta = 1$ corresponds to independence. The Gumbel copula has upper-tail dependence and is useful for datasets where the dependence between high values of the univariate distributions is stronger than the dependence between their low values. The Gumbel copula does not allow negative dependence. Its density is given by

$$c(u, v; \theta) = \frac{C(u, v)}{(uv)} (\log(u)\log(v))^{\theta-1} w^{\frac{2}{\theta}-2} \left(1 + (\theta - 1) w^{\frac{1}{\theta}-2}\right), \tag{S18}$$

where $w = (-log(u))^{\theta} + (-log(v))^{\theta}$. Consequently, the log-likelihood function(Correia Martins André, 2019) is defined by

$$\begin{aligned} &\log\big(L(\theta|u, v)\big) \\ &\quad = w^{\frac{1}{\theta}} - \log(uv) + (\theta - 1)\log\big(\log(\mathrm{u})\log(\mathrm{v})\big) \\ &\quad + \log\left(w^{\frac{2}{\theta}-2}\right)\log\left[(\theta - 1) w^{\frac{1}{\theta}-2}\right]. \end{aligned} \tag{S19}$$

For the bivariate Clayton copula, the generator is $\varphi(t) = \frac{1}{\theta}\left(t^{-\theta} - 1\right)$ and inverse generator $\varphi(s)^{-\theta} = \max\left\{(1 + \theta s)^{-1/\theta}, 0\right\}$ for $\theta \in [-1, \infty)\backslash\{0\}$. The bivariate Clayton copula is given as

$$C(u, v) = max\left(\left(u^{-\theta} + v^{-\theta} - 1\right)^{-\frac{1}{\theta}}, 0\right), \tag{S20}$$

where full dependence is reached as $\theta \to \infty$, while independence is reached as $\theta \to 0$.

The density is given by

$$c(u, v; \theta) = \frac{(\theta + 1)(uv)^{\theta}}{(u^{\theta} v^{\theta} (uv)^{\theta})^{\frac{1}{\theta}+2}} \tag{S21}$$

The log-likelihood function for the Clayton copula is defined by

$$\begin{aligned} &\log\big(L(\theta|u, v)\big) \\ &\quad = \log(1) + \theta - (\theta + 1)(\log(u) + \log(v)) \\ &\quad - \left(\frac{1 + 2\theta}{\theta}\right) \log\big(u^{-\theta} + v^{-\theta} + 1\big) \end{aligned} \tag{S22}$$

The independence copula is a special case of several Archimedian copulas, as well as a special case of the Gaussian copula with a correlation matrix equal to the identity matrix. It has no correlation (copula) parameter and no tail dependence. For the bivariate independence copula, the generator is $\varphi(t) = \exp(-t)$. We applied the independence copula as the special case of the Gumbel copula ($\theta = 1$).

Measuring dependence

We use Kendall's tau to assess dependence between the generation interval and incubation period. Rank correlations such as Kendall's tau only depend on the unique copula of the joint distribution and are therefore invariant to monotone transformations of the marginals (Czado, 2019). The relationship between Kendall's tau and the various copula parameters are listed in Table S5.

Of the three copula that allow for dependence between the parameters (Gaussian, Gumbel, Clayton), only the Gaussian copula considers negative correlation.

Mixture model

We used a Bayesian mixture model to determine the best-fit combination of various copulas and marginal distributions. For all possible combinations of copula and marginal distributions we assign individual likelihoods as the sum of component contributions and formulated the model in terms of latent variables. We considered the four copulas described above, and the gamma, lognormal, and Weibull distributions for the generation interval and incubation period for a total of $M = 4 \times 3 \times 3 = 36$ combinations. These M combinations mix in proportion λ, where $\lambda_m \geq 0$ and $\sum_m^M \lambda_m = 1$.

The outcome is drawn from one of these combinations, the identity of which is controlled by a categorical mixing distribution $z \sim \text{Categorical}(\lambda)$. We fixed Kendall's tau and the means and SDs for the generation interval and incubation period across all possible combinations M. We used informative priors for the means of the generation interval and incubation period, obtained from previous publications (Linton et al., 2020; Nishiura et al., 2020e).

Acknowledgements:

The authors thank the reporting jurisdictions in Japan for conducting thorough investigations and publishing information about cases and transmission settings of public health import, Atsuna Tokumoto for help with initial data collection on transmission links for two prefectures, and the Nishiura lab for additional support with data collection on case characteristics.

Chapter 3 is under review for publication. However, a preprint version has been published on *medRxiv*. Minor formatting modifications and edits have been made for the dissertation.

Conclusion

Nearly two years after SARS-CoV-2 emerged, control remains elusive. The COVID-19 pandemic has proven longer-lasting and more severe than was generally anticipated in early 2020. Partially, this is due to the large fraction of presymptomatic and asymptomatic transmission, which indicated that controlling epidemic spread via isolation of symptomatic cases alone was infeasible (as discussed in *Chapter 3*), and led to the implementation of various other nonpharmaceutical interventions, from border control and lockdowns to mask wearing and physical distancing (Chinazzi et al., 2020; Ferguson et al., 2020). To understand transmission dynamics and fine-tune statistical models of SARS-CoV-2 transmission, COVID-19 severity, and the impact of nonpharmaceutical interventions on control, continued advancements in the estimation and application of the epidemiological parameters related to SARS-CoV-2 infection have been critical.

Chapter 1 presented estimates of the incubation period and other epidemiological parameters of COVID-19 using data reported when the disease had first emerged in humans. Next, *Chapter 2* advanced a previously developed method used to determine when outbreaks end to account for reporting delays and consider underascertainment by estimating the timing of when outbreaks of SARS-CoV-2 would end using clusters reported in Japan. Finally, *Chapter 3* returned to parameter estimation, with the joint estimation of the generation interval and incubation period and assessment of their correlation. From these three chapters, the following insights were produced which contribute to COVID-19 control efforts.

- *Chapter 1:* Estimated the COVID-19 incubation period, which was used to inform quarantine policies, transmission models, and underpinned backcalculation of the time of infection of COVID-19 cases (Abbott et al., 2020; Hamlet et al., 2020). Backcalculation is a critical component used in estimation of SARS-CoV-2 transmission strength in Japan (Jung et al., 2020c).

- *Chapter 1:* Presented estimates of delay distributions from onset to hospitalization, onset to death, and hospitalization to death were used to inform later studies performing similar

estimations or directly used to improve estimates of case or infection fatality risk (Russell et al., 2020; Wilson et al., 2020).

- *Chapter 2:* Investigated end-of-outbreak timing for clusters of COVID-19 cases using a heuristic method. Use of rigorous quantitative methods to determine when outbreaks end and tailoring the parameters and outputs of these methods to address public health and policy goals can release resources and restrictions sooner than untailored estimates and help to ensure that a valid risk assessment has been carried out and that there is only a limited risk of a falsely declaring the end of an outbreak. Issues and concerns related to end-of-outbreak estimation will become more prominent as we come to better define and understand possible COVID-19 “endgame” scenarios.

- *Chapter 2:* Considered when outbreaks end by cluster, rather than at the national or subnational level. Such highly targeted assessments can better direct public health measure relaxation and resource demobilization for small-scale outbreaks such as in medical facilities (Akhmetzhanov et al., 2021) or other places where clusters may occur. This focus on common venues or pathways of pathogen transmission (rather than geography) is important because pathogens do not obey administrative boundaries and not all outbreaks have a clear geographical scope (Coulombier and Takkinen, 2013; Thompson et al., 2016; Wilkinson et al., 2011).

- *Chapter 3:* Assessed whether there were differences in the incubation period and generation interval by subgroup, providing one of the most comprehensive characterizations of the generation interval of wild-type COVID-19. We found that as seen with the serial interval, there was some change in the generation interval over time, which—if not accounted for—could affect estimates of transmission (Ali et al., 2020; Park et al., 2019). In addition, we provided a rare estimate of the generation interval of COVID-19 for asymptomatic infectors.

- *Chapter 3:* Presented the first empirical estimate of correlation between the generation interval and incubation period, showing how this correlation can impact understanding of presymptomatic transmission and isolation-based disease control efforts.

With the emergence of SARS-CoV-2 variants and the rollout of COVID-19 vaccines, control of COVID-19 remains a moving target. However, the present dissertation has provided estimates for key epidemiological parameters for use in the modelling of COVID-19 spread and shown their application to end-of-outbreak estimation as well as to elucidating the role of presymptomatic transmission and effectiveness of symptom-based case isolation. Although there is much work yet to be done to determine epidemiological parameters as they relate to vaccination and SARS-CoV-2 variants, the work presented here will continue to inform COVID-19 statistical models moving forward.

Acknowledgements

While pursing my PhD I received encouragement and support from many, many, people, in addition to financial support from a Ministry of Education, Culture, Sports, Science and Technology (MEXT) Research Scholarship. Here, I express my thanks.

I am eternally grateful for the support of Hiroshi Nishiura, whose depth and breadth of infectious disease modelling expertise propelled my work forward. He provided many opportunities for me throughout my PhD course and has been an inspiring presence during the COVID-19 pandemic, continuously promoting research-informed policy. I can happily conclude my time as a PhD student having learned a great deal under his guidance. An immense thank you as well to Akiko Tamakoshi for supporting my PhD work during the last year and a half of my studies.

I was also fortunate enough to benefit from many conversations and collaborations with Andrei Akhmetzhanov. He encouraged me to think in new ways about various epidemiological problems and approaches to their evaluation and estimation, including while running around the Imperial Palace. Likewise, I am happy to have had the opportunity to work and publish with many wonderful people, including Asami Anzai, Atsuna Tokumoto, Baoyin Yuan, Katsuma Hayashi, Keita Yoshii, Marie Fujimoto, Misaki Sasanami, Robin Thompson, Ryo Kinoshita, Sorano Sumire, Sung-mok Jung, Taishi Kayano, Takeshi Miyama, Tetsuro Kobayashi, and Yichi Yang, among others. I am grateful to have met, collaborated with, and learned from them.

I also want to thank staff I worked with in Sapporo (Keiko Saito, Hisae Hirama, and Miwako Inagi, and Ikumi Karakawa) and Kyoto (Chidori Nishimura, Mari Tatsumi, and Natsuko Matsuura) for supporting my day-to-day activities as a researcher.

My time in Sapporo and Kyoto would not have been the same without the numerous people I have spent days with outdoors in the mountains. I want to thank them for helping to keep me grounded (well, climbing) all over Japan, and leaving me with many happy memories of Japanese mountains in all seasons. Lastly, I am forever grateful to my family, partner, and friends—in the United States and elsewhere—who have met the many international moves I have made in pursuit of professional development as an infectious disease epidemiologist with magnanimous, unwavering support.

Disclosure of conflicts of interest

The author declares no conflict of interest.

References

Abbott, S., Hellewell, J., Thompson, R.N., Sherratt, K., Gibbs, H.P., Bosse, N.I., Munday, J.D., Meakin, S., Doughty, E.L., Chun, J.Y., et al. (2020). Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Res. *5*, 112.

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., and Xia, L. (2020). Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases. Radiology *2019*, 200642.

Akhmetzhanov, A.R., Jung, S., Cheng, H.Y., and Thompson, R.N. (2021). A hospital-related outbreak of SARS-CoV-2 associated with variant Epsilon (B.1.429) in Taiwan: transmission potential and outbreak containment under intensified contact tracing, January–February 2021. Int. J. Infect. Dis. *110*, 15–20.

Ali, S.T., Wang, L., Lau, E.H.Y., Xu, X.-K., Du, Z., Wu, Y., Leung, G.M., and Cowling, B.J. (2020). Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. Science (80-.). *9004*, eabc9004.

Anzai, A., and Nishiura, H. (2021). "Go To Travel" campaign and travel-associated coronavirus disease 2019 cases: A descriptive analysis, July–August 2020. J. Clin. Med.

Backer, J.A., Klinkenberg, D., and Wallinga, J. (2020). Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. Eurosurveillance 1–6.

Baker, M.G., Wilson, N., and Anglemyer, A. (2020). Successful elimination of Covid-19 transmission in New Zealand. N. Engl. J. Med. *383*, e56.

Te Beest, D.E., Wallinga, J., Donker, T., and Van Boven, M. (2013). Estimating the generation interval of influenza a (H1N1) in a range of social settings. Epidemiology *24*, 244–250.

Bi, Q., Wu, Y., Mei, S., Ye, C., Zou, X., Zhang, Z., Liu, X., Wei, L., Truelove, S.A., Zhang, T., et al. (2020). Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. Lancet Infect. Dis. *3099*, 1–9.

Blumberg, S., and Lloyd-smith, J.O. (2013). Inference of R0 and transmission heterogeneity from the size distribution of stuttering chains. PLoS Comput. Biol. *9*, 1–17.

Brauner, J.M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., Stephenson, A.B., Leech, G., Altman, G., Mikulik, V., et al. (2021). Inferring the effectiveness of government interventions against COVID-19. Science (80-.). *802*.

Britton, T., and Tomba, G.S. (2019). Estimation in emerging epidemics: Biases and remedies. J. R. Soc. Interface *16*.

Bushman, M., Worby, C., Chang, H.H., Kraemer, M.U.G., and Hanage, W.P. (2021). Transmission of SARS-CoV-2 before and after symptom onset: Impact of nonpharmaceutical interventions in China. Eur. J. Epidemiol. *36*, 429–439.

Cabinet Secretariat of Japan (2021). COVID-19 emergency declarations.

Chinazzi, M., Davis, J.T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., Piontti, A.P. y, Rossi, L., Sun, K., Viboud, C., et al. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (2019-nCoV) outbreak. Science (80-.). *9757*, 2020.02.09.20021261.

Chinese National Health Commission (NHC) (2021). Experts answer reporters' questions on pneumonia outbreak of new coronavirus infection].

Contreras, S., Dehning, J., Loidolt, M., Zierenberg, J., Spitzner, F.P., Urrea-Quintero, J.H., Mohr, S.B., Wilczek, M., Wibral, M., and Priesemann, V. (2021). The challenges of containing SARS-CoV-2 via test-trace-and-isolate. Nat. Commun. *12*, 1–13.

Cori, A., Ferguson, N.M., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. Am. J. Epidemiol. *178*, 1505–1512.

Correia Martins André, L.M.B. (2019). Copula models for dependence: Comparing classical and Bayesian approaches.

Coulombier, D., and Takkinen, J. (2013). From national to international - Challenges in cross-border multi-country, multi-vehicle foodborne outbreak investigations. Eurosurveillance *18*, 1–2.

Cowling, B.J., and Leung, G.M. (2020). Epidemiological research priorities for public health control of the ongoing global novel coronavirus (2019-nCoV) outbreak. Eurosurveillance *25*, 2000110.

Czado, C. (2019). Analyzing dependent data with vine copulas: A practical guide with R.

Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., Pearson, C.A.B., Russell, T.W., Tully, D.C., Washburne, A.D., et al. (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science (80-.). *372*, eabg3055.

Deng, Y., You, C., Liu, Y., Qin, J., and Zhou, X.H. (2020). Estimation of incubation period and generation time based on observed length-biased epidemic cohort with censoring for COVID-19 outbreak in China. Biometrics 1–13.

Donnelly, C.A., Ghani, A.C., Leung, G.M., Hedley, A.J., Fraser, C., Riley, S., Abu-Raddad, L.J., Ho, L.M., Thach, T.Q., Chau, P., et al. (2003). Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. Lancet *361*, 1761–1766.

Du, Z., Xu, X., Wu, Y., Wang, L., Cowling, B.J., and Meyers, L.A. (2020). Serial interval of COVID-19 among publicly reported confirmed cases. Emerg. Infect. Dis. *26*, 1341–1343.

Endo, A., Abbott, S., Kucharski, A.J., and Funk, S. (2020a). Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome Open Res. 1–8.

Endo, A., Leclerc, Q.J., Knight, G.M., Medley, G.F., Atkins, K.E., Funk, S., and Kucharski, A.J. (2020b). Implication of backward contact tracing in the presence of overdispersed transmission in COVID-19 outbreaks. Wellcome Open Res. *5*, 1–17.

Ferguson, N.M., Laydon, D.J., Nedjati-Gilani, G.L., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z.M., Cuomo-Dannenburg, G., et al. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID- 19 mortality and healthcare demand.

Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., Parker, M., Bonsall, D., and Fraser, C. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science (80-.). *368*, eabb6936.

Fraser, C., Riley, S., Anderson, R.M., and Ferguson, N.M. (2004). Factors that make an infectious disease outbreak controllable. Proc. Natl. Acad. Sci. U. S. A. *101*, 6146–6151.

Furuse, Y., Sando, E., Tsuchiya, N., Miyahara, R., Yasuda, I., Ko, Y.K., Saito, M., Morimoto, K., Imamura, T., Shobugawa, Y., et al. (2020). Clusters of coronavirus disease in communities, Japan, January–April 2020. Emerg. Infect. Dis. *26*, 2176–2179.

Gandhi, M., Yokoe, D.S., and Havlir, D. V. (2020). Asymptomatic transmission, the achilles' heel of current strategies to control Covid-19. N. Engl. J. Med. *382*, 2158–2160.

Ganyani, T., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J., and Hens, N. (2020). Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Eurosurveillance *25*, 1–8.

Gaymard, A., Bosetti, P., Feri, A., Destras, G., Enouf, V., Andronico, A., Burrel, S., Behillil, S., Sauvage, C., Bal, A., et al. (2021). Early assessment of diffusion and possible expansion of SARS-CoV-2 Lineage 20I/501Y.V1 (B.1.1.7, variant of concern 202012/01) in France, January to March 2021. Eurosurveillance *26*, 1–6.

Gostic, K.M., McGough, L., Baskerville, E.B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J.A., De Salazar, P.M., et al. (2020). Practical considerations for measuring the effective reproductive number, Rt. PLoS Comput. Biol. *16*, 1–21.

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., et al. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nat. Hum. Behav. *5*, 529–538.

Hamlet, A., Djafaara, B.A., Cucunubá, Z., Mesa, D.O., Green, W., Thompson, H., Nayagam, S., Ainslie, K.E.C., Bhatia, S., Bhatt, S., et al. (2020). The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. *422*, 413–422.

Hart, W.S., Hochfilzer, L.F.R., Cunniffe, N.J., Lee, H., Nishiura, H., and Thompson, R.N. (2019). Accurate forecasts of the effectiveness of interventions against Ebola may require

models that account for variations in symptoms during infection. Epidemics *29*, 100371.

Hart, W.S., Abbott, S., Endo, A., Hellewell, J., Miller, E., Andrews, N., Maini, P.K., and Thompson, R.N. (2021). Inference of SARS-CoV-2 generation times using UK household data. 1–32.

He, D., Zhao, S., Lin, Q., Zhuang, Z., Cao, P., Wang, M.H., and Yang, L. (2020a). The relative transmissibility of asymptomatic COVID-19 infections among close contacts. Int. J. Infect. Dis. *94*, 145–147.

He, X., Lau, E.H.Y., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y.C., Wong, J.Y., Guan, Y., Tan, X., et al. (2020b). Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat. Med. *26*, 672–675.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., and Gu, X. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet *6736*, 1–10.

Imai, N., Gaythorpe, K.A.M., Abbott, S., Prem, K., Liu, Y., Bhatia, S., Elsland, S. Van, and Ferguson, N.M. (2020a). Adoption and impact of non-pharmaceutical interventions for COVID-19. Wellcome Open Res. *59*, 1–12.

Imai, N., Cori, A., Dorigatti, I., Baguelin, M., Donnelly, C.A., Riley, S., and Ferguson, N.M. (2020b). Report 3: Transmissibility of 2019-nCoV.

Inaba, H., and Nishiura, H. (2008). The basic reproduction number of an infectious disease in a stable population: The impact of population growth rate on the eradication threshold. Math. Model. Nat. Phenom. *3*, 194–228.

Ito, K., Piantham, C., and Nishiura, H. (2021). Predicted domination of variant Delta of SARS-CoV-2 beforeTokyo Olympic Games, Japan, July 2021. Eurosurveillance 4–12.

Jung, S., Kinoshita, R., Thompson, R.N., Linton, N.M., Yang, Y., Akhmetzhanov, A.R., and Nishiura, H. (2020a). Epidemiological identification of a novel pathogen in real time: Analysis of the atypical pneumonia outbreak in Wuhan, China, 2019–2020. J. Clin. Med. *9*, 637.

Jung, S., Akhmetzhanov, A.R., Hayashi, K., Linton, N.M., Yang, Y., Yuan, B., Kobayashi, T., Kinoshita, R., and Nishiura, H. (2020b). Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: Inference using exported cases. J. Clin. Med. *9*, 523.

Jung, S., Akhmetzhanov, A.R., Mizumoto, K., and Nishiura, H. (2020c). Real-time estimation of the effective reproduction number of COVID-19 in Japan.

Kenah, E., Lipsitch, M., and Robins, J.M. (2008). Generation interval contraction and epidemic data analysis. Math. Biosci. *213*, 71–79.

Kissler, S.M., Fauver, J.R., Mack, C., Tai, C.G., Breban, M.I., Watkins, A.E., Samant, R.M., Anderson, D.J., Ho, D.D., Grubaugh, N.D., et al. (2021). Densely sampled viral trajectories

suggest longer duration of acute infection with B.1.1.7 variant relative to non-B.1.1.7 SARS-CoV-2. MedRxiv.

Klinkenberg, D., and Nishiura, H. (2011). The correlation between infectivity and incubation period of measles, estimated from households with two cases. J. Theor. Biol. *284*, 52–60.

Knight, J., and Mishra, S. (2020). Estimating effective reproduction number using generation time versus serial interval, with application to COVID-19 in the Greater Toronto Area, Canada. Infect. Dis. Model. *5*, 889–896.

Knight, G.M., Leclerc, Q.J., and Kucharski, A.J. (2020). Analysis of SARS-CoV-2 transmission clusters and superspreading events.

Kremer, C., Ganyani, T., Chen, D., Torneri, A., Faes, C., Wallinga, J., and Hens, N. (2020). Authors' response: Estimating the generation interval for COVID-19 based on symptom onset data. Eurosurveillance *25*, 18–19.

Lauer, S.A., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., Meredith, H.R., Azman, A.S., Reich, N.G., and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. Ann. Intern. Med. *172*, 577–582.

Leclerc, Q.J., Fuller, N.M., Knight, L.E., Funk, S., and Knight, G.M. (2020). What settings have been linked to SARS-CoV-2 transmission clusters? Wellcome Open Res. *5*, 83.

Lee, H., and Nishiura, H. (2017). Recrudescence of Ebola virus disease outbreak in West Africa, 2014–2016. Int. J. Infect. Dis. *64*, 90–92.

Lee, H., and Nishiura, H. (2019). Sexual transmission and the probability of an end of the Ebola virus disease epidemic. J. Theor. Biol. *471*, 1–12.

Lehtinen, S., Ashcroft, P., and Bonhoeffer, S. (2021). On the relationship between serial interval, infectiousness profile and generation time. J. R. Soc. Interface *18*, 20200756.

Lessler, J., Reich, N.G., and Cummings, D.A.T. (2009). Outbreak of 2009 pandemic influenza A (H1N1) at a New York City School. N. Engl. J. Med. *361*, 2628–2636.

Lessler, J., Reich, N.G., Brookmeyer, R., Perl, T.M., Nelson, K.E., and Cummings, D.A.T. (2015). Incubation periods of acute respiratory viral infections: a systematic review. Lancet Infect. Dis. *9*, 291–300.

Li, M., Liu, K., Song, Y., Wang, M., and Wu, J. (2021). Serial interval and generation interval for imported and local infectors, respectively, estimated using reported contact-tracing data of COVID-19 in China. Front. Public Heal. *8*.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S.M., Lau, E.H.Y., Wong, J.Y., et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N. Engl. J. Med. *382*, 1199–1207.

Linton, N.M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A.R., Jung, S., Yuan, B., Kinoshita, R., and Nishiura, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. J. Clin. Med. *9*.

Liu, Y., Morgenstern, C., Kelly, J., Lowe, R., and Jit, M. (2020). The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. BMC Med. 1–12.

Lloyd-Smith, J.O. (2007). Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PLoS One *2*, 1–8.

Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E., and Getz, W.M. (2005). Superspreading and the effect of individual variation on disease emergence. Nature *438*, 355–359.

Long, D.R., Gombar, S., Hogan, C.A., Greninger, A.L., O'Reilly-Shah, V., Bryson-Cahn, C., Stevens, B., Rustagi, A., Jerome, K.R., Kong, C.S., et al. (2021). Occurrence and timing of subsequent severe acute respiratory syndrome coronavirus 2 reverse-transcription polymerase chain reaction positivity among initially negative patients. Clin. Infect. Dis. *72*, 323–326.

Lovell-Read, F.A., Funk, S., Obolski, U., Donnelly, C.A., and Thompson, R.N. (2021). Interventions targeting nonsymptomatic cases can be important to prevent local outbreaks: SARS-CoV-2 as a case-study. J. R. Soc. Interface *18*.

McAloon, C., Collins, Á., Hunt, K., Barber, A., Byrne, A.W., Butler, F., Casey, M., Griffin, J., Lane, E., McEvoy, D., et al. (2020). Incubation period of COVID-19: A rapid systematic review and meta-analysis of observational research. BMJ Open *10*, 1–9.

Mehand, M.S., Al-shorbaji, F., Millett, P., and Murgue, B. (2018). The WHO R&D Blueprint: 2018 review of emerging infectious diseases requiring urgent research and development effort. Antiviral Res. *159*, 63–67.

Miller, T.E., Garcia Beltran, W.F., Bard, A.Z., Gogakos, T., Anahtar, M.N., Astudillo, M.G., Yang, D., Thierauf, J., Fisch, A.S., Mahowald, G.K., et al. (2020). Clinical sensitivity and interpretation of PCR and serological COVID-19 diagnostics for patients presenting to the hospital. FASEB J. *34*, 13877–13884.

Ministry of Health Labour and Welfare (MHLW) COVID-19 situation and MHLW response (August 10).

Ministry of Health Labour and Welfare (MHLW) (2020). COVID-19 clusters in Japan.

Ministry of Health Labour and Welfare (MHLW) (2021). Response to COVID-19 variants of concern (VOC).

Mizumoto, K., Kagaya, K., Zarebski, A., and Chowell, G. (2020). Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise

ship, Yokohama, Japan, 2020. Eurosurveillance *25*, 1–5.

Nakajo, K., and Nishiura, H. (2021). Transmissibility of asymptomatic COVID-19: Data from Japanese clusters. Int. J. Infect. Dis. *105*, 236–238.

Nishiura, H. (2009). Determination of the appropriate quarantine period following smallpox exposure: An objective approach using the incubation period distribution. Int. J. Hyg. Environ. Health *212*, 97–104.

Nishiura, H. (2010). Time variations in the generation time of an infectious disease: Implications for sampling to appropriately quantify transmission potential. Math. Biosci. Eng. *7*, 851–869.

Nishiura, H., Lee, H.W., Cho, S.H., Lee, W.G., In, T.S., Moon, S.U., Chung, G.T., and Kim, T.S. (2007). Estimates of short- and long-term incubation periods of Plasmodium vivax malaria in the Republic of Korea. Trans. R. Soc. Trop. Med. Hyg. *101*, 338–343.

Nishiura, H., Miyamatsu, Y., and Mizumoto, K. (2016). Objective determination of end of MERS outbreak, South Korea, 2015. Emerg. Infect. Dis. *22*, 146–148.

Nishiura, H., Linton, N.M., and Akhmetzhanov, A.R. (2020a). Initial cluster of novel coronavirus (2019-nCoV) infections in Wuhan, China is consistent with substantial human-to-human transmission. J. Clin. Med. 2019–2021.

Nishiura, H., Kobayashi, T., Miyama, T., Suzuki, A., Jung, S. mok, Hayashi, K., Kinoshita, R., Yang, Y., Yuan, B., Akhmetzhanov, A.R., et al. (2020b). Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). Int. J. Infect. Dis. *94*, 154–155.

Nishiura, H., Jung, S., Linton, N.M., Kinoshita, R., Yang, Y., Hayashi, K., Kobayashi, T., Yuan, B., and Akhmetzhanov, A.R. (2020c). The extent of transmission of novel coronavirus in Wuhan, China, 2020. J. Clin. Med. *9*.

Nishiura, H., Kobayashi, T., Yang, Y., Hayashi, K., Miyama, T., Kinoshita, R., Linton, N.M., Jung, S., Yuan, B., Suzuki, A., et al. (2020d). The rate of underascertainment of novel coronavirus (2019-nCoV) infection: Estimation using Japanese passengers data on evacuation flights. J. Clin. Med. 2019–2021.

Nishiura, H., Linton, N.M., and Akhmetzhanov, A.R. (2020e). Serial interval of novel coronavirus (COVID-19) infections. Int. J. Infect. Dis. 113332.

Omori, R., Mizumoto, K., and Nishiura, H. (2020). Ascertainment rate of novel coronavirus disease (COVID- 19). Int. J. Infect. Dis. *96*, 673–675.

Oshitani, H. (2020). Cluster-based approach to coronavirus disease 2019 (COVID-19) response in Japan, from February to April 2020. Jpn. J. Infect. Dis. *73*, 491–493.

Oshitani, H., and The Expert Members of the National COVID-19 Cluster Taskforce at the Ministry of Health Labour and Welfare of Japan (2020). Cluster-based approach to coronavirus disease 2019 (COVID-19) response in Japan—February–April 2020. Jpn. J. Infect. Dis.

Parag, K. V, Donnelly, C.A., Jha, R., and Thompson, R.N. (2020). An exact method for quantifying the reliability of end-of-epidemic declarations in real time. PLoS Comput. Biol. *16*, e1008478.

Park, M., Cook, A.R., Lim, J.T., Sun, Y., and Dickens, B.L. (2020a). A systematic review of COVID-19 epidemiology based on current evidence. J. Clin. Med. *9*, 967.

Park, S.W., Champredon, D., Weitz, J.S., and Dushoff, J. (2019). A practical generation-interval-based approach to inferring the strength of epidemics from their speed. Epidemics 1–7.

Park, S.W., Bolker, B.M., Champredon, D., Earn, D.J.D., Li, M., Weitz, J.S., Grenfell, B.T., and Dushoff, J. (2020b). Reconciling early-outbreak estimates of the basic reproductive number and its uncertainty: Framework and applications to the novel coronavirus (SARS-CoV-2) outbreak: Reconciling early-outbreak estimates of the basic reproductive number and its uncertainty. J. R. Soc. Interface *17*.

Park, S.W., Cornforth, D.M., Dushoff, J., and Weitz, J.S. (2020c). The time scale of asymptomatic transmission affects estimates of epidemic potential in the COVID-19 outbreak. Epidemics *31*, 100392.

Park, S.W., Champredon, D., and Dushoff, J. (2020d). Inferring generation-interval distributions from contact-tracing data. J. R. Soc. Interface *17*, 20190719.

Park, S.W., Sun, K., Champredon, D., Li, M., Bolker, B., Earn, D., Weitz, J., Grenfell, B.T., and Dushoff, J. (2021). Forward-looking serial intervals correctly link epidemic growth to reproduction numbers. Proc. Natl. Acad. Sci. U. S. A. *118*, e2011548118.

Public Health Agency of Canada (2020). Interim guidance: Public health management of cases and contacts associated with novel coronavirus disease 2019 (COVID-19). Gov. Canada 1–19.

R Development Core Team (2019). R: A Language and Environment for Statistical Computing.

Reich, N.G. (2019). Package 'coarseDataTools.'

Reich, N.G., Lessler, J., Cummings, D.A.T., and Brookmeyer, R. (2009). Estimating incubation period distributions with coarse data. Stat. Med. *28*, 2769–2784.

Riou, J., and Althaus, C.L. (2020). Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. Eurosurveillance *25*, 1–5.

Russell, T.W., Hellewell, J., Jarvis, C.I., Zandvoort, K. Van, Abbott, S., Ratnayake, R., Flasche, S., Eggo, R.M., Edmunds, W.J., and Kucharski, A.J. (2020). Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. Eurosurveillance *25*, 6–10.

Sklar, A. (1959). Fonctions de Répartition à n Dimensions et Leurs Marges. Publ. L'Institut Stat. L'Universit{é} Paris.

Stan Development Team Prior choice recommendations.

Stan Development Team (2021). CmdStan: the command-line interface to Stan.

Tariq, A., Lee, Y., Roosa, K., Blumberg, S., Yan, P., Ma, S., and Chowell, G. (2020). Real-time monitoring the transmission potential of COVID-19 in Singapore, March 2020. BMC Med. *18*, 166.

Thompson, R.N., Cobb, R.C., Gilligan, C.A., and Cunniffe, N.J. (2016). Management of invading pathogens should be informed by epidemiology rather than administrative boundaries. Ecol. Modell. *324*, 28–32.

Thompson, R.N., Morgan, O.W., and Jalava, K. (2019). Rigorous surveillance is necessary for high confidence in end-of-outbreak declarations for Ebola and other infectious diseases. Philos. Trans. R. Soc. B Biol. Sci. *374*, 20180431.

Thompson, R.N., Hollingsworth, T.D., Isham, V., Arribas-Bel, D., Ashby, B., Britton, T., Challenor, P., Chappell, L.H.K., Clapham, H., Cunniffe, N.J., et al. (2020). Key questions for modelling COVID-19 exit strategies. Proceedings. Biol. Sci. *287*, 20201405.

Tindale, L.C., Stockdale, J.E., Coombe, M., Garlock, E.S., Lau, W.Y.V., Saraswat, M., Zhang, L., Chen, D., Wallinga, J., and Colijn, C. (2020). Evidence for transmission of covid-19 prior to symptom onset. Elife *9*, 1–34.

Torneri, A., Libin, P., Scalia Tomba, G., Faes, C., Wood, J.G., and Hens, N. (2021). On realized serial and generation intervals given control measures: The COVID-19 pandemic case. PLOS Comput. Biol. *17*, e1008892.

Virlogeux, V., Park, M., Wu, J.T., and Cowling, B.J. (2016). Association between severity of MERS-CoV infection and incubation period. Emerg. Infect. Dis. *22*, 526–528.

Wilkinson, K., Grant, W.P., Green, L.E., Hunter, S., Jeger, M.J., Lowe, P., Medley, G.F., Mills, P., Phillipson, J., Poppy, G.M., et al. (2011). Infectious diseases of animals and plants: An interdisciplinary approach. Philos. Trans. R. Soc. B Biol. Sci. *366*, 1933–1942.

Wilson, N., Kvalsvig, A., Barnard, L.T., and Baker, M.G. (2020). Case-fatality risk estimates for COVID-19 calculated by using a lag time for fatality. Emerg. Infect. Dis. *26*, 2019–2021.

World Health Organization (WHO) (2020a). Public health criteria to adjust public health and social measures in the context of COVID-19 (12 May 2020).

World Health Organization (WHO) (2020b). WHO recommended criteria for declaring the end of the Ebola virus disease outbreak.

Wu, P., Hao, X., Lau, E.H.Y., Wong, J.Y., Leung, K.S.M., Wu, J.T., Cowling, B.J., and Leung, G.M. (2020). Real-time tentative assessment of the epidemiological characteristics of novel coronavirus infections in Wuhan, China, as at 22 January 2020. Eurosurveillance 1–6.

Xu, X.-K., Liu, X.F., Wu, Y., Ali, S.T., Du, Z., Bosetti, P., Lau, E.H.Y., Cowling, B.J., and Wang, L. (2020). Reconstruction of transmission pairs for novel coronavirus disease 2019 (COVID-19) in Mainland China: Estimation of superspreading events, serial interval, and hazard of infection. Clin. Infect. Dis. *2019*, 1–5.

Yoshioka, T., and Maeda, Y. (2020). Covid-19 stigma induced by local government and media reporting in Japan: It's time to reconsider risk communication lessons from the fukushima daiichi nuclear disaster. J. Epidemiol. *30*, 372–373.

Zhang, J., Litvinova, M., Wang, W., Wang, Y., Deng, X., Chen, X., Li, M., Zheng, W., Yi, L., Chen, X., et al. (2020a). Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. Lancet Infect. Dis. *3099*, 1–10.

Zhang, Y., Li, Y., Wang, L., Li, M., and Zhou, X. (2020b). Evaluating transmission heterogeneity and super-spreading event of COVID-19 in a metropolis of China. Int. J. Environ. Res. Public Health *17*, 3705.