Estimation of the incubation period of COVID-19 using viral load data

Keisuke Ejima [a,1,**], Kwang Su Kim [b,1], Christina Ludema [a], Ana I. Bento [a], Shoya Iwanami [b], Yasuhisa Fujita [b], Hirofumi Ohashi [c,d], Yoshiki Koizumi [e], Koichi Watashi [c,d,f,g], Kazuyuki Aihara [h], Hiroshi Nishiura [i], Shingo Iwami [b,f,j,k,l,*]

[a] Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington, IN, USA
[b] Department of Biology, Faculty of Sciences, Kyushu University, Fukuoka, Japan
[c] Department of Virology II, National Institute of Infectious Diseases, Tokyo, Japan
[d] Department of Applied Biological Science, Tokyo University of Science, Noda, Japan
[e] National Center for Global Health and Medicine, Tokyo, Japan
[f] MIRAI, JST, Saitama, Japan
[g] Institute for Frontier Life and Medical Sciences, Kyoto University, Kyoto, Japan
[h] International Research Center for Neurointelligence, The University of Tokyo, Tokyo, Japan
[i] Graduate School of Medicine, Hokkaido University, Hokkaido, Japan
[j] Institute for the Advanced Study of Human Biology (ASHBi), Kyoto University, Kyoto, Japan
[k] NEXT-Ganken Program, Japanese Foundation for Cancer Research (JFCR), Tokyo, Japan
[l] Science Groove Inc., Fukuoka, Japan

ARTICLE INFO

ABSTRACT

The incubation period, or the time from infection to symptom onset, of COVID-19 has usually been estimated by using data collected through interviews with cases and their contacts. However, this estimation is influenced by uncertainty in the cases' recall of exposure time. We propose a novel method that uses viral load data collected over time since hospitalization, hindcasting the timing of infection with a mathematical model for viral dynamics. As an example, we used reported data on viral load for 30 hospitalized patients from multiple countries (Singapore, China, Germany, and Korea) and estimated the incubation period. The median, 2.5, and 97.5 percentiles of the incubation period were 5.85 days (95 % CI: 5.05, 6.77), 2.65 days (2.04, 3.41), and 12.99 days (9.98, 16.79), respectively, which are comparable to the values estimated in previous studies. Using viral load to estimate the incubation period might be a useful approach, especially when it is impractical to directly observe the infection event.

1. Introduction

The current COVID-19 outbreak is characterized by a longer incubation period (i.e., time from infection to symptom onset) than that of influenza and other acute respiratory viruses. This longer incubation period means that many of the strategies for disease control that rely on symptom-based surveillance (e.g., community fever monitoring or home observation of travelers for symptoms) will not effectively control the outbreak. For example, the wide geographic spread of SARS-CoV-2 could have been driven by this long incubation period, allowing cases to pass through border control measures such as temperature screening

(Wells et al., 2020).

Estimating the incubation period is challenging, because we rarely directly observe the time of infection or the time of symptom onset (examples to the contrary in HIV infection show the intense follow-up needed to observe these events (Robb et al., 2016; Rolland et al., 2020)). The first study estimating the incubation period of SARS-CoV-2 was that of Bi et al. (2020), who fit a log-normal model to a subset of cases for whom detailed information about exposure to another case was available.

However, even with meticulous contact tracing, directly observing infector-infectee pairs is a time-consuming process, especially when the

---

* Corresponding author at: 744, Motooka, Nishi-ku, Fukuoka, Japan 819-0395.
** Corresponding author at: PH158A, 1025 E 7th St, Bloomington, IN, USA 47405.
E-mail addresses: kejima@iu.edu (K. Ejima), iwamishingo@gmail.com (S. Iwami).
[1] These authors contributed equally to this work.

incubation period is lengthy. Measuring the incubation period through contact tracing is more difficult if the infector-infectee pair had a lot of contact with each other, leading to a number of suspected individuals needing to be interviewed. Indeed, Bi et al. demonstrated large uncertainty (the interval of exposure was more than 10 days for about 25 % of the cases) concerning the timing of infection for COVID-19 in China (Bi et al., 2020). Although a majority of studies estimated incubation period of SARS-CoV-2 (Bi et al., 2020; Lauer et al., 2020; Linton et al., 2020) used a statistical modelling technique that accounts for uncertainty in both the reports of exposure time and the time of symptom onset (Reich et al., 2009), they had to inherently use a heuristic weight function for the censored information.

Here we propose another approach to estimating the incubation period, in which we use longitudinal data on viral load and hindcast the point of initial infection. Viral load data were collected at the early stage of the epidemic for clinical purposes (e.g., understanding the etiology and the pathophysiology of COVID-19) and to ensure patients were no longer shedding virus (or more precisely, viral fragments) before hospital discharge. The data were analyzed using a mathematical model describing viral dynamics, which typically draws a bell-shaped curve (i. e., viral load first increases exponentially until the peak, where it starts

to decline). Although the data are available only after the onset of symptoms, the timing of infection can be estimated by hindcasting the model for each case.

2. Results

2.1. Viral load dynamics for SARS-CoV-2

We extracted viral load data for 30 hospitalized patients as reported in four papers (**Table S1**) and quantified the dynamics of SARS-CoV-2 infection with a mathematical model previously proposed (Ikeda et al., 2016; Perelson, 2002; Kim et al., 2020a):

$$\frac{df(t)}{dt} = -\beta f(t)V(t), \quad \frac{dV(t)}{dt} = \gamma f(t)V(t) - \delta V(t),$$

where $f(t)$ and $V(t)$ are the relative fraction of uninfected target cells at time t to those at time 0 and the amount of virus at time t, respectively. The parameters β, γ, and δ are the rate constant for virus infection, the maximum rate constant for viral replication, and the death rate of infected cells, respectively. The viral load data from the four different
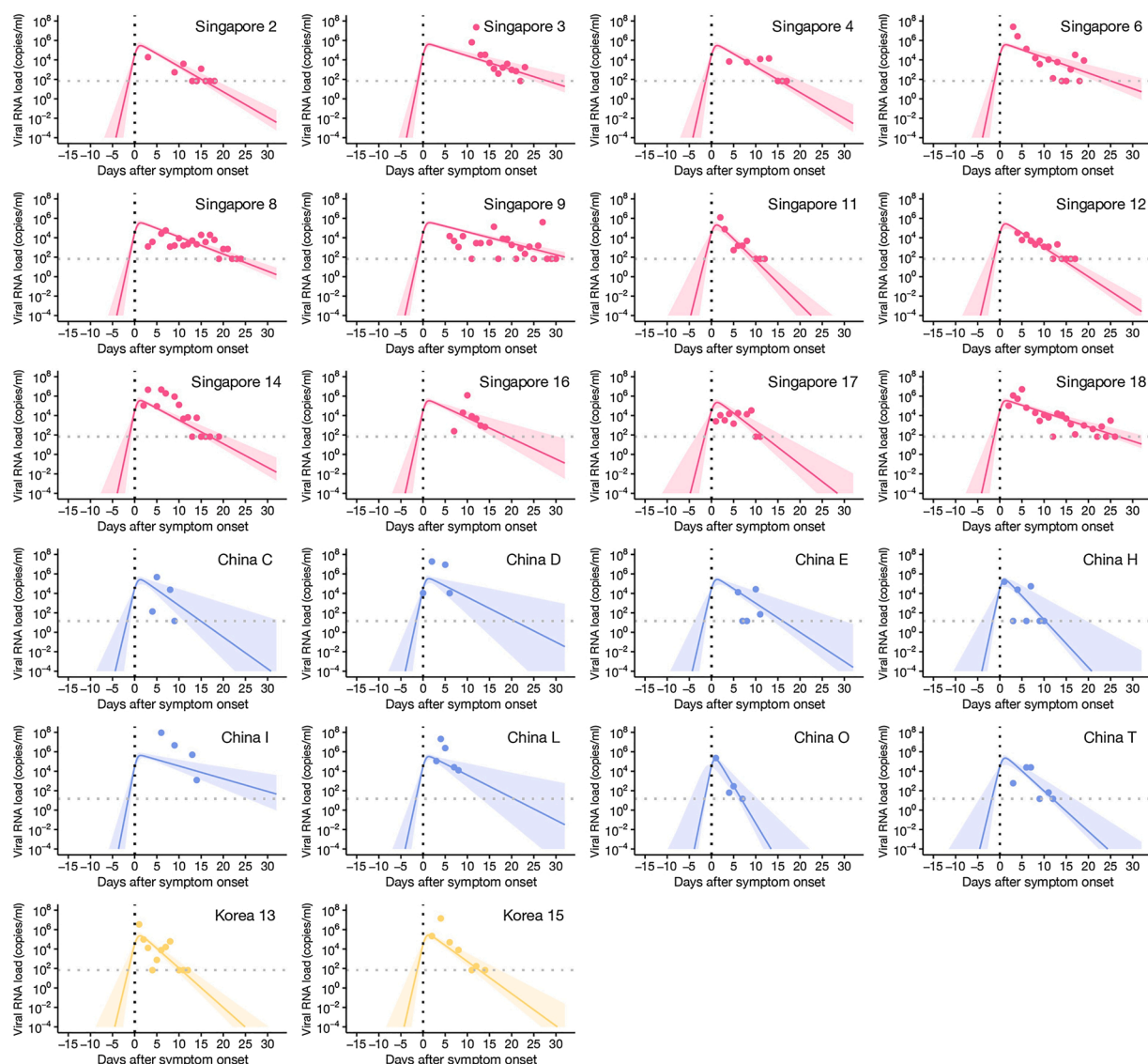

Fig. 1. Viral load dynamics for each case in Asia: Each colored symbol corresponds to the measured viral load (Singapore: pink, China: blue, Korea: yellow). The shadowed region corresponds to the estimated viral load from 100 sets of parameters resampled from conditional distributions; the solid line gives the best-fit curve. The time scale is days since the onset of symptoms (the black dotted vertical line is the day of symptom onset). The gray dashed horizontal line is the detection limit.

papers were fitted to the model with mixed effects, which assumed that the parameters for each individual follow normal distributions with the same population mean. Several different models other than the model described above are available to explain the viral load trajectory of acute infection. However, we chose this model because it better explains the data. As an example, the model considering eclipse phase has been proposed for acute infection and has been applied to SARS-CoV-2 (Baccam et al., 2006; Gonçalves et al., 2020). We fitted the models with and without eclipse phase and compared the goodness-of-fit (i.e., BIC). Although the BICs were comparable, we needed to fix the parameter value that determines the length of the eclipse phase. Thus, we decided to use the current model without eclipse phase. The viral load dynamics for each case in Asia and Europe is shown in Figs. 1 and 2, respectively, and the estimated values of the parameters for each case are summarized in **Table S2**. The peak of viral load appeared 2–3 days after symptom onset. Note that in the data, there were no cases in which viral load was measured before symptom onset. Among the total 30 cases, viral load was the highest in the first measurement in 14 cases.

2.2. Establishing a viral load threshold for infection

To assess the day on which SARS-CoV-2 infection was established, in other words, the start of the exponential growth phase of the viral load (Perelson, 2002), we needed to set the viral load threshold for this timing. The time of the infection event, T_{inf}, was identified by means of back-calculation by using the dataset when the viral load reaches the threshold. We used the three cases reported from China (Patients D, H, and L) with known primary cases to determine the viral load threshold to establish infection (Zou et al., 2020). For these three cases, the day of exposure was assumed to be equal to the day of the infection event, as follows. A case (Patient E) from Wuhan visited Patient D and Patient L in Zhuhai on January 17. Patients D and L developed symptoms on January 23 and 20, respectively (thus their primary case is Patient E). Two cases (Patient I and P) from Wuhan visited their daughter, Patient H, in Zhuhai on January 11. Patient H developed fever on January 17 (thus her primary cases are Patient I and P). Using this contact information, we computed the viral load for the day on which infection was established by hindcasting the mathematical model with the estimated parameters, which we defined as the *infection establishment threshold*. The viral load threshold was $10^{-6.59} - 10^{-5.01}$, $10^{-6.49} - 10^{-4.93}$, and $10^{-1.77} - 10^{-0.20}$ (all of which correspond to 95 % confidence intervals) for Patients D, H, and L, respectively. The viral load threshold, or the dose of exposure, should be heterogeneous between patients; thus, we fitted the normal distribution to the log-transformed estimated viral load threshold. Specifically, we randomly sampled 100 values from the estimated distribution of viral load threshold for each of the three patients and then fitted a normal distribution. We used this distribution ($10^x; x \sim N(-4.17, 2.25^2)$) as the threshold for further analyses. Because the viral load thresholds estimated for the three patients

differed substantially, we performed the same analysis for each patient and used the thresholds to estimate the distribution of incubation periods as sensitivity analyses.

2.3. Incubation period of COVID-19

With the viral load threshold, we computed the incubation period, T_{inf}, for all patients by hindcasting the mathematical model after fitting the model to the data. To address the uncertainty of the estimation, we resampled 100 parameter sets for each individual including the viral load threshold and obtained the corresponding 100 T_{inf} for each individual (i.e., 100×30 T_{inf} in total) (see "The nonlinear mixed effect model" for the details of computation). Then, the three parametric distributions were fitted to 100×30 T_{inf}: Weibull, gamma, and log-normal distributions. Comparing the Akaike Information Criteria (AIC) for those three distributions, the best model (i.e., that with the lowest AIC) was used for further analyses. The parametric bootstrap method was used to assess parameter uncertainty. Specifically, the bootstrap sample was composed of 30 T_{inf}: a single T_{inf} was resampled from the 100 T_{inf} of each individual. The best parametric model (i.e., Weibull, gamma, or log-normal distribution) was fitted to the bootstrapped data for parameter inference. We repeated this process 1000 times and obtained 1000 parameter sets, and the median, 2.5, and 97.5 percentiles of the distribution were computed. As a sensitivity analysis, the above process was repeated with the data from Europe (Germany) and Asia (China, Singapore, and Korea) separately.

Fig. 3 shows the 95 % CI of the empirical distribution of the timing of infection for each case estimated from the viral load data using the virus dynamics model. The AICs of the three models (log-normal, gamma, and Weibull distributions) were 13812.7, 13930.3, and 14390.1, respectively. Thus, the lognormal distribution was preferred and was used for further analysis. Fig. 4**A** and **D** summarize the estimated cumulative distribution function and probability density function of the incubation period for COVID-19 (log-normal distribution), respectively. The median, 2.5, and 97.5 percentiles of the incubation period were 5.85 days (95 % CI: 5.05, 6.77), 2.65 days (2.04, 3.41), and 12.99 days (9.98, 16.79), respectively. For a sensitivity analysis, distributions of the incubation period were computed for Asia (Singapore, China, and Korea) and Europe (Germany) separately in Fig. 4**BC** and **EF**. The median, 2.5, and 97.5 percentiles of the incubation period were 5.77 days (95 % CI: 4.81, 6.61), 2.69 days (1.84, 3.49), and 12.26 days (9.30, 16.67), respectively, for Asian countries. The median, 2.5, and 97.5 percentiles of the incubation period were 6.01 days (95 %CI: 4.93, 7.37), 3.01 days (1.76, 4.13), and 12.18 days (7.98, 16.91), respectively, for Europe. We did not observe large difference in the incubation period in the data from Asia and Europe. Furthermore, we used the threshold estimated from each patient to estimate the distribution of the incubation period. The median, 2.5, and 97.5 percentiles of the incubation period were 5.81 days (95 % CI: 4.97, 6.69), 2.53 days (1.96, 3.25), and 13.23 days
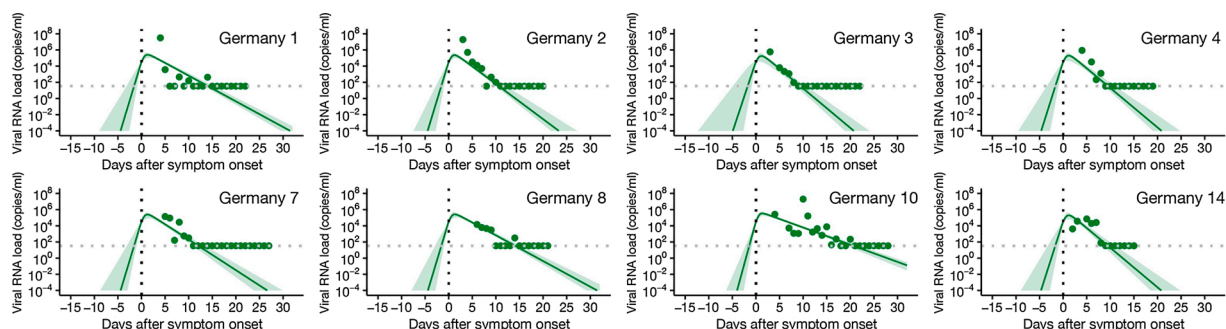

Fig. 2. Viral load dynamics for each case in Europe: Each colored symbol corresponds to the measured viral load (Germany: green). The shadowed region corresponds to the estimated viral load from 100 sets of parameters resampled from conditional distributions; the solid line gives the best-fit curve. The time scale is days since the onset of symptoms (the black dotted vertical line is the day of symptom onset). The gray dashed horizontal line is the detection limit.
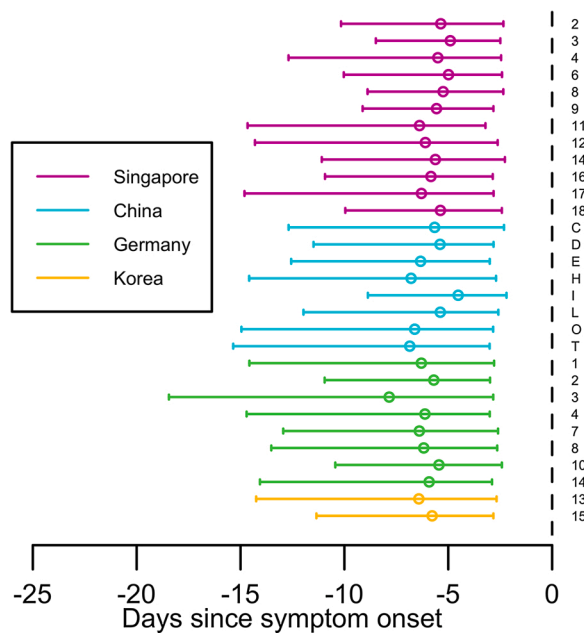
Fig. 3. The estimated day on which infection was established for each case by use of day from symptom onset as the time scale: The dots and the bars are the median, 2.5, and 97.5 percentiles of the empirical distribution of the estimated day on which infection was established for each case. The case IDs on the right correspond to those in the original papers.

(10.46, 17.23) using the threshold of Patient D; 7.17 days (95 % CI: 6.37, 8.02), 3.85 days (3.21, 4.65), and 13.35 days (10.74, 16.55) using the threshold of Patient H; and 3.89 days (95 % CI: 3.45, 4.37), 2.04 days

(1.72, 2.48), and 7.29 days (5.81, 9.02) using the threshold of Patient L. As expected, the lower threshold yielded to a longer incubation period.

3. Discussion

Inferring the timing of infection is challenging in general. Given asymptomatic and presymptomatic transmission and the relatively long incubation period of SARS-CoV-2, not all patients are aware of how they were exposed or the specific time of exposure. The median incubation period of SARS-CoV-2 is estimated to be 5–6 days (Bi et al., 2020; Lauer et al., 2020; Linton et al., 2020; Backer et al., 2020), whereas that for other acute respiratory viral infections, such as SARS-CoV-1, non-SARS human coronaviruses, influenza A virus, and influenza B virus, are estimated to be 4.0, 3.2, 1.4, and 0.6 days, respectively (Lessler et al., 2009). The proportion of SARS-CoV-2 infection that is asymptomatic ranges from 40 % to 45 % (Prevalence of Asymptomatic SARS-CoV-2 Infection), which is close to that for influenza (50 %) (Weinstein et al., 2003). By contrast, asymptomatic cases are rarely observed for SARS-CoV-1 (Lee et al., 2003). Thus, we proposed using viral load data, which are externally measured and are independent of recall. The median of the estimated incubation period was about 6 days, and 97.5 % of cases developed symptoms in about 13 days. These estimations are consistent with previously published estimates (Bi et al., 2020; Lauer et al., 2020; Linton et al., 2020; Backer et al., 2020).

Mass vaccination campaigns for SARS-CoV-2 have been proceeding with unprecedented speed; however, the risk of resurgence will not be negligible (influenza outbreaks happen even though effective vaccines are available). In addition to the vaccine, contact tracing is important to reduce the risk for resurgence, and being able to make valid estimates of the incubation period helps to reduce the burden in the contact tracing process. Indeed, contact tracing helped to further identify and treat cases earlier than a symptom-based approach (Bi et al., 2020). Furthermore,
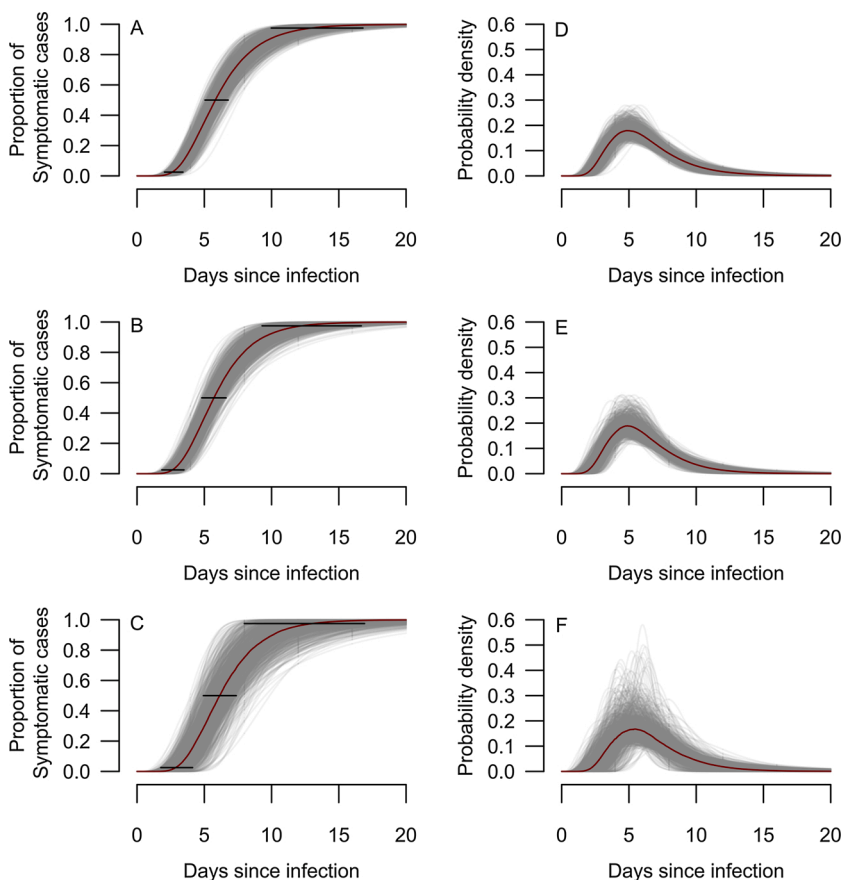

Fig. 4. The estimated incubation period: (A, B, C) The cumulative distribution function for total, Asian, and European cases, respectively. We used the log-normal distribution for fitting. The gray lines were drawn based on the 1000 different bootstrap samples. The horizontal bars are 95 % CIs at 2.5 %, 50 %, and 97.5 % of the distribution. The solid red curve corresponds to the median of the estimated distribution. (D, E, F) The probability density function for total, Asian, and European cases, respectively.

when we know the incubation period distribution, we can better assess the role of presymptomatic infection in the outbreak. Combined with the serial-interval distribution, the incubation-period distribution has been used to quantify the proportion of presymptomatic infection (He et al., 2020).

The strength of this approach is that it can complement the limitations of the classic interview-based approach regarding ascertaining the exposure event. Our proposed approach may be applicable not only to human infectious disease and zoonoses such as influenza and COVID-19, but also to animal/livestock infectious diseases such as foot-and-mouth disease when contact recall is not possible. Furthermore, replicating viral load from infection to recovery is helpful not only for estimating the incubation period but also for clinical and epidemiologic understanding of the disease. For example, we observed that the viral load of SARS-CoV-2 peaked 2–3 days after the onset of symptoms, which is consistent with the finding that the viral load in throat swabs was on the decline when first measured (2–4 days since symptom onset) (18, 19).

There are several limitations to be noted in this study. One is related to the modelling approach. Our approach did not account for any uncertainty in reported day of symptom onset because the data did not include the range of exposed days. The approach accounting for uncertainty was previously proposed by Reich et al. (Reich et al., 2009). Combining our approach with that of Reich et al. might reduce uncertainty surrounding the precise reporting of exposure and illness onset once such data are available, which is doable because estimation of the timing of infection and estimation of the incubation-period distribution are independent. The model we used in this study did not include detailed immune response or antiviral effects given limited information. We can update the model once relevant data are available.

Another limitation is relevant to the data we used. The proposed approach requires collection of viral loads over time since symptom onset, which might not be feasible for all patients or in resource-limited contexts. A few studies have investigated change in viral load over time (Benefield et al., 2020; Kucirka et al., 2020). However, those studies included viral load data with observation at a single time point because they did not consider individual variability. Further, one paper assumed that the incubation period was 5 days when symptom onset information was not available (Benefield et al., 2020). We believe such an approach is unreasonable because *1)* the day of exposure is extremely hard to observe (and therefore we are proposing to use longitudinal viral load data), and *2)* the incubation period varies between patients. We admit the inclusion criterion for data in our study (more than three data points from each patient) is a limitation of our study; however, we do not think that adding nonlongitudinal data or data without information on symptom onset would be an option. We used data from hospitalized and symptomatic patients. If viral dynamics and the incubation period differ in unhospitalized patients, the estimated incubation-period distribution should represent that for hospitalized patients only. Indeed, we are planning to collect saliva samples from mildly symptomatic to asymptomatic patients (https://rctportal.niph.go.jp/en/detail?trial_id=jRCT2071200023). We used the viral load data collected from upper respiratory specimens (i.e., nasopharyngeal, oropharyngeal, nasal swabs), because viral dynamics differs between organs, as evidenced in multiple studies (e.g., rectal swab vs. nasal swab) (Xu et al., 2020; Young et al., 2020). However, viral dynamics might also differ between nasal, nasopharyngeal, and oropharyngeal swabs, even though they are close. Similarly, sex, age, and other factors might influence viral dynamics; however, such information was not consistently available from all patients. We treated the different types of swabs as a covariate in the model, but the computation did not converge because of the small sample sizes. However, because we used a mixed-effect model, the random effect in every parameter (on each patient) should have considered the difference in viral dynamics due to the sample type and demographic differences to some extent.

Being able to make valid estimates of the incubation period distribution is essential for mitigating risk of disease spread. Knowing the

estimated incubation period distribution simplifies the process of contact tracing and improves our understanding of the role of presymptomatic infection. By unifying the proposed approach with existing epidemiologic methods, we can achieve precise estimation of the incubation-period distribution.

4. Materials and methods

4.1. Data

The viral load data were from 30 hospitalized patients presented in four previously published studies of hospitalized COVID-19 patients (Zou et al., 2020; Wölfel et al., 2020; Young et al., 2020; Kim et al., 2020b). All cases used in our analysis presented with symptoms before or after hospitalization. For consistency, the viral load data from upper respiratory specimens were used in the analysis. Patients treated with antivirals or with less than two data points were excluded. For all the studies from which we extracted data, ethics approval was obtained from the ethics committee at each institute. Written informed consent was obtained from the patients or their next of kin in the original studies. We summarized the data in **Table S1** and described the details in the **Supplementary Material**.

4.2. The nonlinear mixed effect model

MONOLIX 2019R2 (www.lixoft.com) was used to fit the nonlinear mixed-effects model to the viral load data. The nonlinear mixed-effects model incorporates fixed effects and random effects accounting for interpatient variability in viral dynamics. Specifically, the parameter for individual k, $\theta_k (= \theta \times e^{\pi_k})$, is represented by the fixed effect, θ, and the random effect, π_k, which follows the Gaussian distribution with mean 0 and standard deviation Ω. The fixed effect (population parameter) and random effect were estimated by using the stochastic approximation EM (expectation-maximization) algorithm and empirical Bayes' method, respectively. Using estimated parameters and a Markov Chain Monte Carlo algorithm, we obtained the conditional distribution of parameters for each patient. A total of 100 parameter sets for each patient were resampled from the conditional distribution and used to estimate the incubation period distribution.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.epidem.2021.100454.

References

Baccam, P., Beauchemin, C., Macken, C.A., Hayden, F.G., Perelson, A.S., 2006. Kinetics of influenza a virus infection in humans. J. Virol. 80 (15), 7590–7599.

Backer, J.A., Klinkenberg, D., Wallinga, J., 2020. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20-28 January 2020. Euro Surveill. 25 (5).

Benefield, A.E., et al., 2020. SARS-CoV-2 viral load peaks prior to symaptom onset: a systematic review and individual-pooled analysis of coronavirus viral load from 66 studies. medRxive 2020, 2009.2028.20202028.

Bi, Q., 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. 20 (8), 911–919.

Gonçalves, A., et al., 2020. Timing of antiviral treatment initiation is critical to reduce SARS-CoV-2 viral load. CPT Pharmacometrics Syst. Pharmacol. 9 (9), 509–514.

He, X., et al., 2020. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat. Med. 26 (5), 672–675.

Ikeda, H., et al., 2016. Quantifying the effect of Vpu on the promotion of HIV-1 replication in the humanized mouse model. Retrovirology 13 (1), 23.

Kim, K.S., et al., 2020a. Modelling SARS-CoV-2 dynamics: implications for therapy. medRxiv 2020, 2003.2023.20040493.

Kim, E.S., et al., 2020b. Clinical course and outcomes of patients with severe acute respiratory syndrome coronavirus 2 infection: a preliminary report of the first 28 patients from the korean cohort study on COVID-19. J. Korean Med. Sci. 35 (13), e142.

Kucirka, L.M., Lauer, S.A., Laeyendecker, O., Boon, D., Lessler, J., 2020. Variation in false-negative rate of reverse transcriptase polymerase chain reaction-based SARS-CoV-2 tests by time since exposure. Ann. Intern. Med. 173 (4), 262–267.

Lauer, S.A., et al., 2020. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. Ann. Intern. Med. 172 (9), 577–582.

Lee, H.K.K., et al., 2003. Asymptomatic severe acute respiratory syndrome-associated coronavirus infection. Emerging Infect. Dis. 9 (11), 1491–1492.

Lessler, J., et al., 2009. Incubation periods of acute respiratory viral infections: a systematic review. Lancet Infect. Dis. 9 (5), 291–300.

Linton, N.M., et al., 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 (2), 538.

Perelson, A.S., 2002. Modelling viral and immune system dynamics. Nature reviews. Immunology 2 (1), 28–36.

Prevalence of Asymptomatic SARS-CoV-2 Infection (https://www.acpjournals.org/doi/10.7326/M20-3012).

Reich, N.G., Lessler, J., Cummings, D.A.T., Brookmeyer, R., 2009. Estimating incubation period distributions with coarse data. Stat. Med. 28 (22), 2769–2784.

Robb, M.L., et al., 2016. Prospective study of acute HIV-1 infection in adults in East Africa and Thailand. N. Engl. J. Med. 374 (22), 2120–2130.

Rolland, M., et al., 2020. Molecular dating and viral load growth rates suggested that the eclipse phase lasted about a week in HIV-1 infected adults in East Africa and Thailand. PLoS Pathog. 16 (2), e1008179.

Weinstein, R.A., Bridges, C.B., Kuehnert, M.J., Hall, C.B., 2003. Transmission of influenza: implications for control in health care settings. Clin. Infect. Dis. 37 (8), 1094–1101.

Wells, C.R., et al., 2020. Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. Proc. Natl. Acad. Sci. 117 (13), 7504–7509.

Wölfel, R., et al., 2020. Virological assessment of hospitalized patients with COVID-2019. Nature 581 (7809), 465–469.

Xu, Y., et al., 2020. Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. Nat. Med. 26 (4), 502–505.

Young, B.E., et al., 2020. Epidemiologic features and clinical course of patients infected with SARS-CoV-2 in Singapore. JAMA 323 (15), 1488–1494.

Zou, L., et al., 2020. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. N. Engl. J. Med. 382 (12), 1177–1179.