# Estimation of the percentages of undiagnosed patients of the novel coronavirus (SARS-CoV-2) infection in Hokkaido, Japan by using birth-death process with recursive full tracing

Takuma Tanaka[1,2*], Takayuki Yamaguchi[2], Yohei Sakamoto[3]

**1** Graduate School of Data Science, Shiga University, Hikone, Shiga, Japan
**2** The Center for Data Science Education and Research, Shiga University, Hikone, Shiga, Japan
**3** Graduate School of Medicine, The Jikei University School of Medicine, Minato, Tokyo, Japan

* tanaka.takuma@gmail.com

## Abstract

Estimating the percentages of undiagnosed and asymptomatic patients is essential for controlling the outbreak of SARS-CoV-2, and for assessing any strategy for controlling the disease. In this paper, we propose a novel analysis based on the birth-death process with recursive full tracing. We estimated the numbers of undiagnosed symptomatic patients and the lower bound of the number of total infected individuals per diagnosed patient before and after the declaration of the state of emergency in Hokkaido, Japan. The median of the estimated number of undiagnosed symptomatic patients per diagnosed patient decreased from 1.7 to 0.77 after the declaration, and the median of the estimated lower bound of the number of total infected individuals per diagnosed patient decreased from 4.2 to 2.4. We will discuss the limitations and possible expansions of the model.

## Introduction

The novel coronavirus (SARS-CoV-2) spread to the most populated areas of the world in the first few months of 2020. In Japan, the first case was reported on January 16th, 2020; on March 31st, the number of cases increased to 2122 [1]. In Hokkaido, the largest prefecture of Japan, the first case was reported on January 28th, 2020 (Fig 1). The Hokkaido government declared a state of emergency on February 28th and lifted it on March 19th. The state of emergency was not legally binding, and the government asked the residents to stay home on the weekends. Until the state of emergency was lifted, a total of 157 cases were reported. Out of 157, two died, and eight have recently traveled abroad. Although the number of diagnosed cases is declining towards the end of March, the situation remains uncertain.

To effectively control the spread of the infection, we need to know several parameter values characterizing the infection, such as the basic reproduction number, $R_0$, the percentage of asymptomatic patients, and the fatality rate. One of the factors that complicate the decision making in disease control is the uncertainty in the percentage of asymptomatic patients. Several lines of evidence indicate that the virus can be transmitted by asymptomatic patients [2, 3], who may have facilitated the

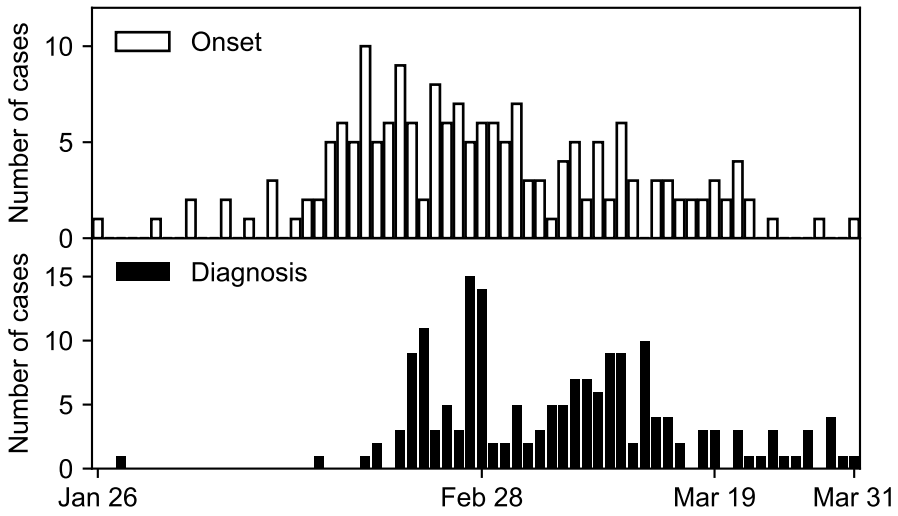


**Fig 1. Epidemic curves of COVID-19 in Hokkaido, early 2020.** The histograms of the date of onset (above) and the date of diagnosis (below) are shown.

spread of SARS-CoV-2. To make the disease control strategy successful and less harmful to the society, the transmissibility of asymptomatic patients and the percentage of asymptomatic patients should be measured. Another complicating factor is the uncertainty in the percentages of diagnosed and undiagnosed patients. Most individuals infected with SARS-CoV-2 suffer from mild cold-like symptoms and recover without any medical intervention. They remain undiagnosed, disturbing how we monitor the number of total infected individuals. Thus, estimating the percentages of asymptomatic and undiagnosed patients is a challenging problem to be answered in epidemiology.

It would be useful if these values could be estimated by the information available in the early phase of the outbreak. Contact tracing is considered to be one of the effective measures, and health officials have conducted contact tracing of infected patients to prevent the spread of the virus infection for outbreaks of new or reemerging infections [4–9]. If an infected patient is found, health officials try to find other infected people among those who have come into contact with the infected patient. If any other infected patients are found, officials will repeat tracing from the newly found patients; and if not, tracing is stopped. Thus, a "cluster" of infected patients connected by the route of infection is constructed through the process of contact tracing. Simultaneously, the dates of onset and diagnosis are obtained for each diagnosed patient. As a result, contact tracing can provide detailed qualitative and quantitative information about diagnosed patients from the infectious disease transmission network of diagnosed and undiagnosed patients. Analyzing this network with an appropriate model may enable us to estimate the parameter values of the infection.

One promising model for contact tracing is a stochastic model based on the birth-death process, which is a formulation of branching processes [10–17], because the number of cases in the early phase stochastically fluctuates and the widely used deterministic SIR models are inapplicable. In birth-death processes, a sequence of infectious events generates a tree whose nodes are infected patients and edges are infection routes (Fig 2). When a patient recovers (Fig 2, dotted circle), the

corresponding node and its edges are removed from the tree, which is split into two. The infectious disease transmission network is composed of trees. A connected component of the network is referred to as "cluster" in this paper. Contact tracing corresponds to finding a node in the tree and removing nodes connected to the first node found (Fig 2, gray filled circle). There are various types of contact tracing based on how to choose nodes to be removed, including backward, forward, or full tracing, and recursive or one-step tracing [17]. We consider only recursive full tracing, that is, all nodes directly and indirectly connected to the first found node are removed (Fig 2, dashed rounded rectangle). We propose an analysis based on the birth-death process with recursive full tracing that takes advantage of information obtained by contact tracing to estimate epidemiological parameter values with a small set of data. We focus on the contact tracing of infected individuals in Hokkaido. The present analysis uses the distributions of the cluster size and patients' time from onset to diagnosis, which are released by the health officials, to estimate the model parameters. Our approach directly models the stochastic dynamics, which is an inherent property of the early phase of the outbreak.

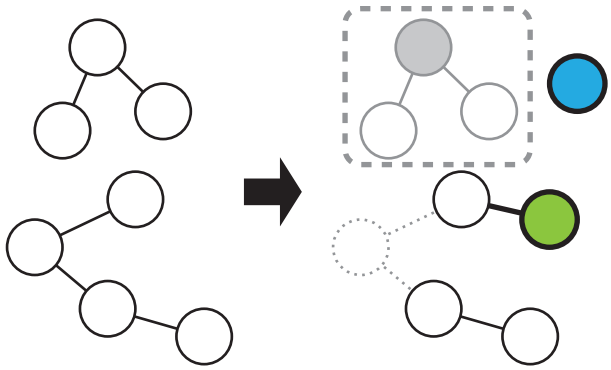

**Fig 2. Infection, recovery, and diagnosis in the birth-death process with recursive full tracing.** A network with two clusters is shown on the left side, and the network after the progression of events is shown on the right side. The nodes represent symptomatic patients, and two nodes are connected by an edge if one has infected the other. The nodes can recover and be removed from the network (dotted circle/lines), or infect and connect to a new node (green circle). A new node without edges can be generated in the network (blue circle). A node can be diagnosed (gray filled circle), and the nodes in the same connected component (gray open circles) are removed from the network and counted among diagnosed symptomatic clusters (dashed rounded rectangle). Nodes and edges removed from the network are indicated in gray, and those newly generated in the network are indicated by bold lines.

This paper is organized as follows. In the Methods section, we summarize the SARS-CoV-2 infection in Hokkaido and divide it into those before and after the declaration of the state of emergency. We classify patients into symptomatic and asymptomatic, and diagnosed and undiagnosed. We explain what corresponds to diagnosed and undiagnosed patients in the present model. We describe the formulation of the model and the details of the simulations. In the Results section, we estimate the parameter values and the number of asymptomatic and undiagnosed patients. In the Discussion section, we relate our results with previous studies and discuss the limitations and possible expansions of the model.

## Methods

This paper reports the analysis of the SARS-CoV-2 infection in Hokkaido, Japan [18]. In Hokkaido, all cases had not traveled abroad recently except for three cases, which include a tourist from Wuhan, China. We concentrate our analysis on the cases whose onset was prior to the lifting of the state of emergency. We excluded the cases with an unknown onset and asymptomatic patients from the analysis. If the date of the diagnosis of a case was not reported, it was assumed to coincide with the date of the announcement. S1 Table summarizes the case reports released by the Novel Coronavirus Response Headquarters of the Hokkaido government until April 2nd, 2020.

We represent the patients with nodes and their contacts with edges in a network. If two patients were in close contact with each other, the corresponding nodes are connected by an edge. The network consists of distinct connected components, which we refer to as clusters. Sporadic patients are regarded as size-1 clusters. Fig 3 shows the clusters with sizes larger than 2 in Hokkaido. There are 79 size-1 clusters and 20 size-2 clusters along with the clusters shown in Fig 3 (S1 Table).

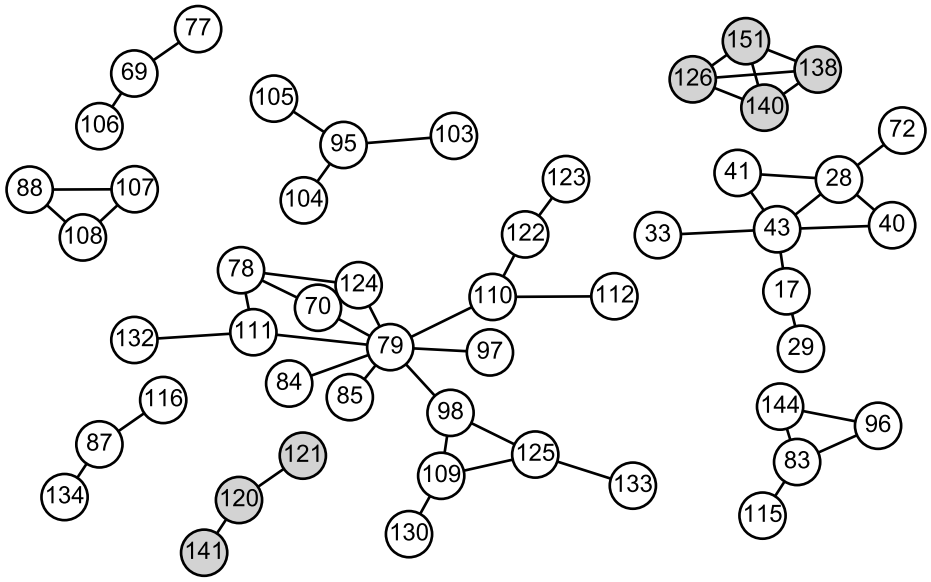


**Fig 3. Contact network between patients in Hokkaido.** The white and gray circles represent the patients in dataset 1 and 2, respectively. The numbers in the circles are the case IDs. Two circles are connected by an edge if these two patients were in close contact with each other. Only the clusters with sizes larger than 2 are shown. There are 59 size-1 clusters and 12 size-2 clusters in dataset 1 and 20 size-1 clusters and 8 size-2 clusters in dataset 2.

All cases were divided into datasets 1 and 2 according to the cluster they belong to. If the earliest onset of the cases in a cluster was prior to the declaration of the state of emergency, this cluster was included in dataset 1; if the earliest onset was between the declaration and the lifting thereof, it was included in dataset 2. Datasets 1 and 2 contain 78 and 30 clusters, respectively. Because the declaration of the state of emergency might have changed the behavior of residents and health officials in Hokkaido, we compared the data before the declaration, dataset 1, and the data between the declaration and the lifting thereof, dataset 2. Table 1 summarizes the datasets.

The patients included in datasets 1 and 2 were all diagnosed and mostly

**Table 1. Summary of datasets.**

| | Dataset 1 | Dataset 2 | Dataset 1 & 2 |
|---|---|---|---|
| Patients | 126 | 43 | 169 |
| Clusters | 78 | 30 | 108 |
| Average time from onset to diagnosis of the patient diagnosed first in a cluster | 9.3 | 6.6 | 8.5 |
| Average time from onset to diagnosis of the all diagnosed patient | 8.4 | 6.2 | 7.8 |
| Average cluster size | 1.6 | 1.4 | 1.6 |
| Largest degree | 8 | 3 | 8 |
| Average degree | 0.87 | 0.74 | 0.84 |
| Average clustering coefficient | 0.070 | 0.093 | 0.076 |

symptomatic. However, not all of the individuals infected with SARS-CoV-2 were diagnosed and symptomatic; they can be classified into diagnosed symptomatic, diagnosed asymptomatic, undiagnosed symptomatic, and undiagnosed asymptomatic groups. The diagnosed symptomatic group consists of those who developed symptoms and were diagnosed, or those who were found in contact tracing. All individuals belonging to this group were covered by our datasets. Although the diagnosed asymptomatic group was also included in the datasets, we ignored this group because this group included only two individuals. The undiagnosed symptomatic group is comprised of individuals who were infected and developed symptoms, but recovered or died without being diagnosed. This group is not directly observable, and thus its percentage is one of the parameters we tried to estimate with the model, which takes this group into account. The undiagnosed asymptomatic group is not directly observable either. It has been suggested that a percentage of SARS-CoV-2 carriers do not develop symptoms but can infect others [2, 19]. We did not explicitly incorporate the asymptomatic group into the model but estimated its percentage.

The birth-death processes have been used to model infectious diseases and population dynamics [10, 11, 16, 17]. The birth-death process is a continuous-time Markov process in which the state variable increases and decreases by one. The increase (birth) and decrease (death) of the state variable occur at the rates which depend on the state variable. In the case of infectious diseases, the state variable is the number of infected individuals, and the birth and death rates are the infection and recovery rates positively correlated with the state variable. The birth-death process can be regarded as a discretized SIS model if the birth and death rates are appropriately set. However, some modifications are needed to model the infection dynamics SARS-CoV-2 in Hokkaido. Since only a fraction of infected individuals are diagnosed and found in the contact tracing, we do not know the exact number of infected individuals. We have to estimate the number of undiagnosed individuals by using the information obtained in the contact tracing. The process of diagnosis and contact tracing must be explicitly included in the model to take advantage of information on the contact between infected individuals.

We modeled the contact tracing of SARS-CoV-2 with a variant of the continuous-time birth-death process, referred to as the birth-death process with recursive full tracing [17]. We used an extended model [10, 13] to take into account the effect of the stochastic diagnosis of individuals and the elimination of the cluster from the population. The model consists of a network whose nodes and edges are continually generated and removed (Fig 2). The lifetime of a node is a random variable drawn from the exponential distribution with the scale parameter $1/\gamma$. During its lifetime, a node gives birth to nodes according to a Poisson point process with the

stationary rate $\beta'$ and is connected to these offspring nodes (Fig 2, green circle). When the lifetime of a node ends, the node and its edges are removed from the network (Fig 2, dotted circle). For the infection dynamics of SARS-CoV-2, the nodes and their birth and death can be regarded as symptomatic patients and their onset of symptoms or infection and recovery from the disease. In the limit of an infinite number of nodes, the dynamics of the number of nodes of birth-death processes are approximated by the SIR model. Asymptomatic infected individuals are not included in this model, and the incubation period is ignored. The difference between the model by Müller and colleagues [10] and the present model is as follows. First, in the former, contact tracing can not necessarily find all infected individuals in the cluster of a diagnosed patient whereas, in the latter, contact tracing finds all individuals in the cluster. Second, the influx of infected individuals is incorporated only in the present model.

The birth-death process with recursive full tracing incorporates the diagnosis and quarantine of patients in addition to the features of continuous-time birth-death processes. The time from infection to diagnosis of a node is a random variable drawn from the exponential distribution with the scale parameter $1/\kappa$. If the diagnosis occurs earlier than the recovery, the node is removed from the network at the diagnosis, representing the quarantine of the diagnosed patient (Fig 2, gray filled circle). At the same time, the nodes in the connected component containing the diagnosed node are also removed from the network, which corresponds to the contact tracing of the infected individuals (Fig 2, gray open circles). Infections that are to be caused by these removed nodes later than the removal are abolished because the diagnosed individuals are quarantined. The connected component of the diagnosed individuals corresponds to the clusters in the datasets (Fig 2, dashed rounded rectangle). The nodes in the connected component are counted among diagnosed symptomatic groups. The recovery of a node disconnects its neighboring nodes. For example, if the green node in Fig 2 is diagnosed, the cluster of size 2 but not of size 4 is reported. A node is not diagnosed if it has been already removed. If a node recovers before it is diagnosed, this node is counted among undiagnosed symptomatic groups.

The simulation of the model is implemented as follows. Nodes without edges are generated in the network according to a Poisson point process with the stationary rate $\lambda = 10^{-5}$ (Fig 2, blue circle). The value of $\lambda$ is inconsequential if it is small enough to allow for observing the cluster size and time from onset to diagnosis distribution at the steady state. On the generation of node $i$ at time $t_i^0$, its time of recovery $t_i^r$ and time of diagnosis $t_i^d$ are assigned as

$$t_i^r - t_i^0 \sim \text{Exponential}(\gamma), \tag{1}$$

$$t_i^d - t_i^0 \sim \text{Exponential}(\kappa), \tag{2}$$

where the recovery rate $\gamma$ and the diagnosis rate $\kappa$ are positive constants. During $t_i^0 \leq t \leq \min(t_i^r,\ t_i^d)$, node $i$ generates new nodes and connects to them according to a Poission point process with the stationary rate $\beta' > 0$. At $t = t_i^r$, if node $i$ is present in the network, node $i$ and its edges are removed from the network. At $t = t_i^d$, if node $i$ is present in the network, the nodes in the same connected component containing node $i$ and their edges are removed from the network.

Let us note that $\beta'$ is the rate of infection giving rise to symptomatic patients, not the rate of infection giving rise to symptomatic and asymptomatic patients. This is because all nodes in the model are capable of being diagnosed, which is not the case with asymptomatic individuals. Hence, $\beta$, the rate of infection giving rise to any type of patient is greater than $\beta'$. For $\kappa = 0$, the probability that a node directly infects $n$ nodes, that is, the probability that a symptomatic patient directly infects $n$

symptomatic patients, follows

$$p(n) = \int_0^\infty \frac{(\beta' T)^n}{n!} \exp(-\beta' T)\gamma \exp(-\gamma T) \mathrm{d}T$$
$$= \frac{\gamma}{\beta' + \gamma} \left( \frac{\beta'}{\beta' + \gamma} \right)^n, \tag{3}$$

whose expected value is $\beta'/\gamma$. Similarly, the basic reproduction number $R_0$ is given by $\beta/\gamma$. Thus, $\beta$ must be greater than $\gamma$ because the number of reported cases is steadily increasing, that is,

$$R_0 = \beta/\gamma > 1. \tag{4}$$

Throughout the simulations reported in this paper, $\gamma$ was fixed to 1/14 [20–22].

We performed the approximate Bayesian computation of the posterior distribution of $\kappa$ and $\beta'$ given the average cluster size and the average time from onset to diagnosis. We drew $\beta'$ from U(0.001, 0.2) and $\kappa$ from U(0.001, 0.12) and accepted the parameter sets with which the average cluster size was identical to 126/78 for dataset 1 and 43/30 for dataset 2, and the average time from onset to diagnosis lied within ±1 days of 9.3 for dataset 1 and 6.6 for dataset 2. The cluster size is defined as the number of nodes of the cluster. The average time from onset to diagnosis of clusters is defined as the average of $t_i^d - t_i^0$ where $i$ runs over the nodes that are diagnosed first in the cluster.

We chose these two summary statistics, the average cluster size and the average time from onset to diagnosis of clusters, to fit the model parameters because of the following reasons. First, these can be obtained without using sophisticated techniques. Second, these allow the precise determination of $\kappa$ and $\beta'$ (see the Results section). Although the number of diagnosed individuals in a given period may be used as a summary statistic, since the number of diagnosed individuals depends on the influx of infected individuals, i.e., the rate of generation of nodes without edge, the parameter value of $\lambda$ as well as $\kappa$ and $\beta'$ must be estimated in this case. In contrast, because the average cluster size and the average time from onset to diagnosis of clusters do not depend on $\lambda$ if $\lambda$ is small enough, $\kappa$ and $\beta'$ can be more precisely estimated by using these two summary statistics. Third, although the parameter value of the birth-death process can be estimated by the time course of the number of infected individuals, the Hokkaido government has not reported the date of recovery of cases, and the time course of the number of infected individuals is unavailable. These two summary statistics can be calculated without information on the recovery of infected individuals. Fourth, although the higher order moments of the cluster size distribution and the time from onset to diagnosis of clusters can be used, they provide almost identical information to their average. Hence, we used the averages because an average is generally more robust than higher order moments. Likewise, other network statistics were not used.

In each run of the simulation, we removed $C_0 + C$ clusters of diagnosed symptomatic patients, of which the first $C_0 = 100$ clusters were discarded to eliminate the dependence of the results on the initial condition and used the following $C = 78$ (dataset 1) and $C = 30$ (dataset 2) clusters as the simulated clusters of patients in Hokkaido. This procedure is justified by the fact that the average cluster size of birth-death processes converges to its steady-state value on a timescale of $1/\beta$ and $1/\gamma$ [11]. This fact also suggests that the properties of clusters in the early phase of the spreading of SARS-CoV-2 can be described by the steady-state of the model.

The ratio of undiagnosed symptomatic patients to diagnosed symptomatic patients can be estimated by the number of nodes that recover without being diagnosed in a period divided by the number of diagnosed nodes that recover in the same period. We used the period between the removal of the $C_0$-th cluster and the removal of the

$C_0 + C$-th cluster to calculate the ratio. This period is referred to as the target period in the following. The ratio of the number of symptomatic and asymptomatic patients to the number of symptomatic patients is $\beta/\beta'$. Because $\beta > \gamma$, which follows from Eq 4, and $\beta > \beta'$, the lower bound of the number of all infected individuals that recover in the target period is estimated by the number of diagnosed and undiagnosed symptomatic patients that recover in the period multiplied by $\max(\gamma,\ \beta')/\beta'$. The estimates presented in this paper are rounded to two significant digits. S2 code is the code that was used to analyse the datasets and generate the figures.

# Results

We performed simulations with randomly generated 100 000 parameter sets and accepted the parameter sets that replicated the average cluster size and the average time from onset to diagnosis of clusters. Before applying the parameter estimation to datasets 1 and 2, we tested whether the present model can successfully estimate the parameter values of an artificial data. The artificial data was generated by a simulation run of the model with $\beta' = 0.1$ and $\kappa = 0.05$. In this simulation, the average cluster size was 2.1, and the average time from onset to diagnosis of clusters was 5.6. The orange crosses, blue triangles, and filled circles in Fig 4A indicate the parameter sets of the simulation runs that replicated the average cluster size, the average time from onset to diagnosis of clusters, and both together, respectively. Out of 100 000 parameter sets, 123 replicated the both summary statistics. This figure shows that the filled circles are concentrated on the intersection of the bands of crosses and triangles. The 95% credible intervals (C.I.) for $\beta'$ and $\kappa$ were $[0.079,\ 0.15]$ and $[0.031,\ 0.085]$, respectively, which successfully contain the parameter set that was used in the original simulation.

Figs 4B and 4C present the parameter sets that replicated the average cluster size and the average time from onset to diagnosis of clusters of datasets 1 and 2, respectively. The 95% C.I. for $\beta'$ and $\kappa$ were $[0.033,\ 0.067]$ and $[0.012,\ 0.042]$ for dataset 1, and $[0.027,\ 0.090]$ and $[0.027,\ 0.11]$ for dataset 2. In dataset 1, the median of the estimated value of $\kappa$, 0.025, was far less than $\gamma$, suggesting that most of the symptomatic patients were not diagnosed before their recovery. The median of the estimated value of $\kappa$ in dataset 2, 0.063, implies that a larger percentage of symptomatic patients were diagnosed in dataset 2.

To examine the number of undiagnosed symptomatic patients, we calculated the number of nodes that recovered without being diagnosed in the target period. Fig 5A and C show the number of undiagnosed symptomatic patients per diagnosed symptomatic patient for dataset 1 and 2, respectively. The 95% C.I. were $[0.95,\ 3.8]$ (median 1.7) for dataset 1 and $[0.35,\ 1.9]$ (median 0.77) for dataset 2. The lower bound of the number of all infected individuals is estimated by the number of diagnosed and undiagnosed symptomatic patients multiplied by $\max(\gamma,\ \beta')/\beta'$ (Fig 5B, D). The 95% C.I. of the lower bound of the total number of infected individuals per diagnosed patient were $[2.3,\ 8.2]$ (median 4.2) for dataset 1 and $[1.5,\ 6.3]$ (median 2.4) for dataset 2, the former of which is consistent with a previous estimate, 1/0.14 [23]. These estimates suggest that around half of infected individuals remain asymptomatic. This is consistent with a previous report on a cruise ship, in which 334 out of 712 infected individuals remained asymptomatic [24]. The 95% C.I. of the lower bound of the total numbers of infected individuals who recovered before the declaration of the state of emergency and those who recovered between the declaration and lifting were $[290,\ 1000]$ (median 530) and $[64,\ 270]$ (median 110), respectively.

To examine the sensitivity of the estimates on the value of $\gamma$, we performed the simulations with $\gamma = 1/10$ and $\gamma = 1/20$. The medians of the number of undiagnosed

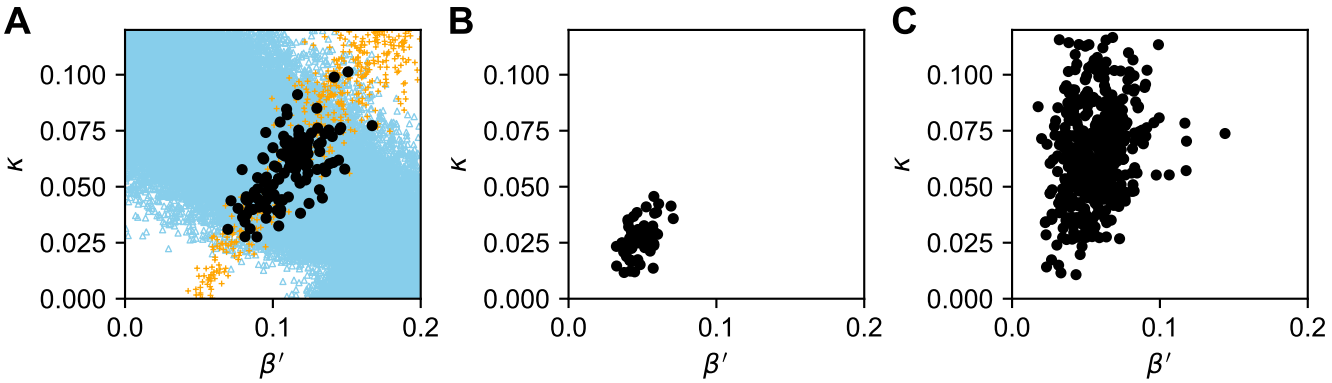


**Fig 4. The estimated parameter sets for the simulated data and datasets 1 and 2.** The parameter sets that replicated both of the average cluster size and the average time from onset to diagnosis of clusters of a simulation run with $\beta' = 0.1$ and $\kappa = 0.05$ (A), dataset 1 (B), and dataset 2 (C) are indicated by the filled circles. Panel A also shows the parameter sets that replicated the average cluster size (orange crosses) and the average time from onset to diagnosis of clusters (blue triangles).

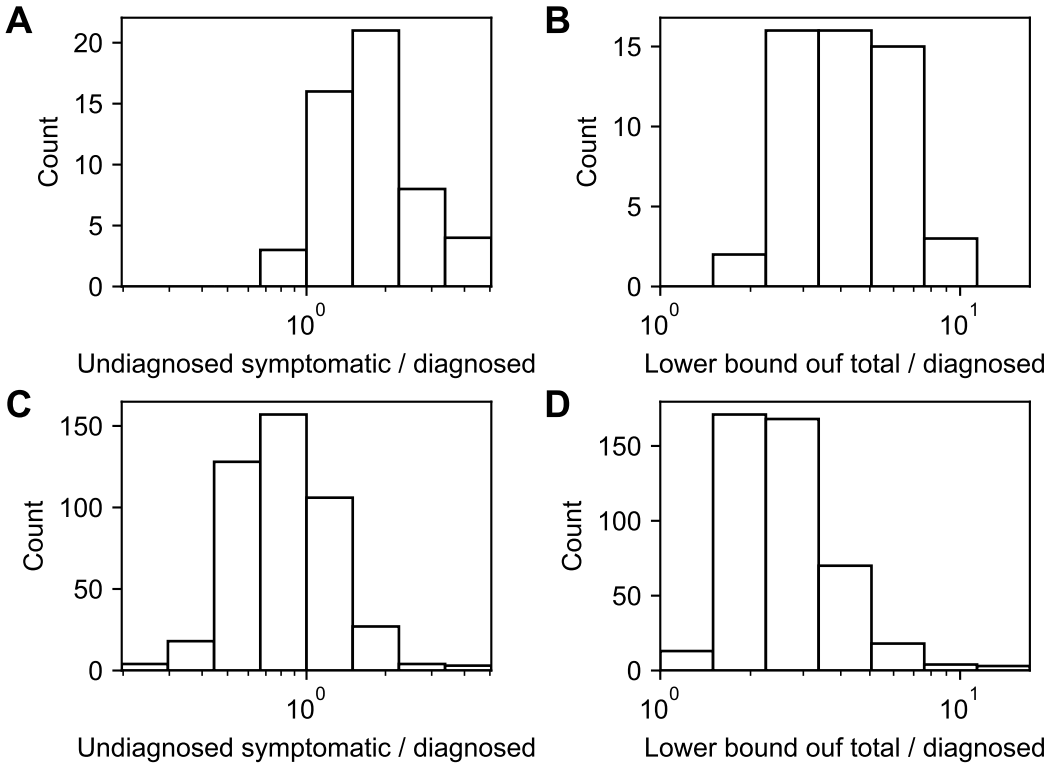


**Fig 5. The estimated number of undiagnosed symptomatic patients (A, C) and the estimated lower bound of the total number of patients (B, D) per diagnosed symptomatic patient. Panels A and B are for dataset 1, and panels C and D are for dataset 2.**

symptomatic patients per diagnosed symptomatic patient for dataset 1 and 2 were 5.8 and 1.5 for $\gamma = 1/10$ and 0.79 and 0.45 for $\gamma = 1/20$, respectively. The medians of the

lower bound of the total number of infected individuals per diagnosed patient for dataset 1 and 2 were 8.9 and 3.4 for $\gamma = 1/10$ and 2.9 and 2.0 for $\gamma = 1/20$, respectively. The medians of the lower bound of the total numbers of infected individuals who recovered before the declaration of the state of emergency and those who recovered between the declaration and lifting were 1100 and 150 for $\gamma = 1/10$ and 370 and 88 for $\gamma = 1/20$, respectively. Hence, a greater $\gamma$, i.e., a shorter mean infective period increases the estimated number of undiagnosed patients, and a less $\gamma$, i.e., a longer mean infective period decreases the estimated number of undiagnosed patients.

## Discussion

In this paper, we have formulated a model to describe the spreading of infection and the quarantine of infected individuals, and estimated the number of undiagnosed symptomatic and asymptomatic COVID-19 patients in Hokkaido. The estimated percentages of undiagnosed symptomatic and asymptomatic patients coincided with previous studies [23, 25]. The estimated lower bound of the total number of patients that recovered before the declaration of a state of emergency was also consistent with a previous study, which estimated that the cumulative incidence in Hokkaido was 2297 cases on February 27th [26]. One of the previous studies approximated the time evolution of the number of infected individuals with differential equations [23], while another estimated the number of asymptomatic patients by using RT-PCR (reverse transcription polymerase chain reaction) test results of evacuees from Wuhan, China on chartered flights [25]. The present analysis focuses on the stochastic dynamics of a discrete number of infected individuals. Thus, the size distribution of clusters, which is a piece of information available in the early phase of the pandemic but difficult to use in differential-equation-based models, can be utilized by the model. Although the methods of the previous reports and ours are completely different, quantitative agreement between them suggests the effectiveness of these approaches.

There are several reasons we have chosen the cases in Hokkaido as the subject of this paper. Hokkaido is an island isolated from the other regions of Japan. In other words, we can assume that a relatively small percentage of the population commutes between Hokkaido and other parts of the world. This makes Hokkaido an ideal subject of the investigation. Until March 20th, one day after the lifting of the state of emergency, 1549 out of 1707 individuals tested with RT-PCR turned out to be negative for SARS-CoV-2 [27], indicating that extensive contact tracing was performed. On the other hand, among the 158 cases diagnosed until March 19th in Hokkaido, only two were asymptomatic. This suggests that in most contact tracing in Hokkaido, RT-PCR tests for SARS-CoV-2 were conducted only on symptomatic patients because of restricted resources. We excluded the diagnosed asymptomatic patients from the analysis because they comprise less than 2% of the reported cases. If the individuals who came into contact with diagnosed patients had been so intensively tested that a much larger number of asymptomatic patients had been diagnosed, the analysis would have needed an extended model including asymptomatic diagnosed patients. The criterion of inclusion and exclusion of asymptomatic individuals in RT-PCR tests could have changed over time, but this is beyond the scope of the present study.

The claim by the local government that test capability was strengthened after the declaration of the state of emergency [28] is supported by a larger value of $\kappa$ in dataset 2 than in dataset 1. The Hokkaido government started reporting the number of RT-PCR tests on March 3rd, four days after the declaration. Until March 3rd, 79 out of 604 were positive (13%), and, from March 4th to March 20th, 79 out of 1103 were positive (7.1%). Although the number of RT-PCR tests before March 3rd is not available, this is consistent with the strengthening of the test capability. Compared to

$\kappa$, $\beta'$ did not exhibit a great change before and after the declaration. This seems counterintuitive because the declaration should have changed the behavior of residents and lowered $\beta'$. There are several possible explanations for this puzzle. Each cluster is classified into datasets 1 and 2 depending on the earliest onset of the cases in the cluster. Because an incubation period precedes the onset of symptoms, some of the clusters in dataset 2 might reflect the spread of SARS-CoV-2 before the declaration, blurring the difference between datasets 1 and 2. A much larger dataset may reveal the change of $\beta'$. Another explanation is that, because the state of emergency was not legally binding, the behavioral modification of the residents was not sufficient to reduce $\beta'$; in contrast, the declaration urged the health officials to strengthen the test capability. Discerning these possibilities needs further study.

Although the Hokkaido datasets were an ideal subject for the present model, the model is not necessarily suitable for other datasets. Kyushu and Shikoku, both of which are islands in Japan, contain several prefectures and, consequently, several local governments. Because the cases reported by these local governments must be merged, the analysis of the spread of SARS-CoV-2 in these islands is much more difficult than that in Hokkaido. If a nation-wide dataset was available, it might be an ideal subject for the present model because of the recent travel restrictions. However, the present model cannot be applied to the dataset that exhibits the superspreader phenomenon because the number of individuals directly infected by an individual follows a geometric distribution (Eq 3). While the cluster size in Hokkaido was moderate as shown in Fig 3, there were large clusters in other prefectures in Japan, for example, the cluster of a club with live music in Osaka [29]. Moreover, our model cannot utilize information on asymptomatic diagnosed patients.

One of the features of the present model is its simplicity. The model has only three essential parameters. The simplicity of the model allowed us to estimate the number of asymptomatic and undiagnosed patients without using a large number of parameter values estimated by previous studies. In the early phase of the spreading of infectious diseases, this simple model can enable the estimation of the asymptomatic and undiagnosed patients despite limited data. Our approach can estimate the number of patients without using costly and time-consuming techniques such as RT-PCR.

Also, its simplicity might allow for an analytical solution. The model is an extension of the birth-death processes, which has been studied intensively. The birth-death processes with contact tracing is analytically tractable [10, 11, 16, 17]. Because the model is simple enough, the analytical solution of the cluster size distribution and the expected time from onset to diagnosis of clusters might be obtained. In future work, the analytical solution would enable us to efficiently estimate the parameter values.

The simplicity of the present model allows expansions in several ways. First, we assumed that $\beta'$ is a fixed value in this paper. This is justified by Fig 3, which shows that the largest number of individuals infected by an individual, that is, the largest degree of nodes, is eight, which is rather small. However, $\beta'$ may be heterogeneous. Heterogeneity in $\beta'$, which has been suggested for the coronavirus genus [30, 31], may explain the superspreader phenomenon. Also, the spread of SARS-CoV-2 might be more accurately modeled with the contact process on scale-free networks [32, 33]. The value of $\beta'$ might depend on the severity of the patient [23]. Second, the recovery rate $\gamma$ may depend on the severity of the patient. Heterogeneity in $\gamma$ can affect the estimated number of total infected individuals. Third, the incubation period, which is ignored in the present paper, might affect the size and structure of clusters [34]. Infectiousness in the incubation period should be included in the model [34]. Fourth, the stage of symptoms should be introduced into the model. Fig 6 suggests that the time from onset to diagnosis of clusters obeys a unimodal distribution with a peak at

around 10 days, although the peak must be at 0 in the present model. Assuming that a mildly infected state stochastically develops into a severely infected state would explain this time course. Fifth, recursive full tracing might be unrealistic because some of the symptomatic patients can be missed in contact tracing. Introducing stochasticity into contact tracing can enable a more precise modeling of clusters. These extensions would be useful in monitoring and controlling the spread of SARS-CoV-2.

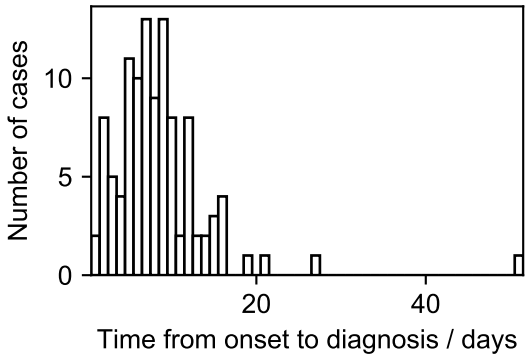


**Fig 6. Distribution of the time from onset to diagnosis of clusters in dataset 1 and 2.**

## Supporting information

**S1 Table. The cases of COVID-19 in Hokkaido, Japan.**

**S2 Code. The code used to analyze the data.**

## Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number JP19K19429.

## References

1. Ministry of Health, Labour, and Welfare, Japan. Press release on April 1st, 2020 [in Japanese]; accessed 4 April 2020. Available from: `https://www.mhlw.go.jp/stf/newpage_10645.html`.
2. Bai Y, Yao L, Wei T, Tian F, Jin DY, Chen L, et al. Presumed asymptomatic carrier transmission of COVID-19. JAMA. 2020;doi:10.1001/jama.2020.2565.
3. Li C, Ji F, Wang L, Wang L, Hao J, Dai M, et al. Asymptomatic and human-to-human transmission of SARS-CoV-2 in a 2-family cluster, Xuzhou, China. Emerging Infectious Disease journal. 2020;26(7). doi:10.3201/eid2607.200718.
4. Donnelly CA, Ghani AC, Leung GM, Hedley AJ, Fraser C, Riley S, et al. Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. The Lancet. 2003;361(9371):1761–1766. doi:10.1016/S0140-6736(03)13410-1.

5. Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-Raddad LJ, Hedley AJ, et al. Transmission Dynamics of the Etiological Agent of SARS in Hong Kong: Impact of Public Health Interventions. Science. 2003;300(5627):1961–1966. doi:10.1126/science.1086478.

6. Saurabh S, Prateek S. Role of contact tracing in containing the 2014 Ebola outbreak: a review. African health sciences. 2017;17(1):225–236. doi:10.4314/ahs.v17i1.28.

7. Eames KTD, Keeling MJ. Contact tracing and disease control. Proceedings of the Royal Society of London Series B: Biological Sciences. 2003;270(1533):2565–2571. doi:10.1098/rspb.2003.2554.

8. House T, Keeling MJ. The impact of contact tracing in clustered populations. PLOS Computational Biology. 2010;6(3):1–9. doi:10.1371/journal.pcbi.1000721.

9. Kwok KO, Tang A, Wei VWI, Park WH, Yeoh EK, Riley S. Epidemic models of contact tracing: systematic review of transmission studies of severe acute respiratory syndrome and middle east respiratory syndrome. Computational and Structural Biotechnology Journal. 2019;17:186–194. doi:10.1016/j.csbj.2019.01.003.

10. Müller J, Kretzschmar M, Dietz K. Contact tracing in stochastic and deterministic epidemic models. Mathematical Biosciences. 2000;164(1):39–64. doi:10.1016/S0025-5564(99)00061-9.

11. Müller J, Möhle M. Family trees of continuous-time birth-and-death processes. Journal of Applied Probability. 2003;40(4):980–994. doi:10.1239/jap/1067436095.

12. Klinkenberg D, Fraser C, Heesterbeek H. The effectiveness of contact tracing in emerging epidemics. PLOS ONE. 2006;1(1):1–7. doi:10.1371/journal.pone.0000012.

13. Müller J, Hösel V. Estimating the tracing probability from contact history at the onset of an epidemic. Mathematical Population Studies. 2007;14(4):211–236. doi:10.1080/08898480701612857.

14. Ball FG, Knock ES, O'Neill PD. Threshold behaviour of emerging epidemics featuring contact tracing. Advances in Applied Probability. 2011;43(4):1048–1065. doi:10.1239/aap/1324045698.

15. Ball FG, Knock ES, O'Neill PD. Stochastic epidemic models featuring contact tracing with delays. Mathematical Biosciences. 2015;266:23–35. doi:10.1016/j.mbs.2015.05.007.

16. Müller J, Koopmann B. The effect of delay on contact tracing. Mathematical Biosciences. 2016;282:204–214. doi:10.1016/j.mbs.2016.10.010.

17. Okolie A, Müller J. Exact and approximate formulas for contact tracing on random trees. Mathematical Biosciences. 2020;321:108320. doi:10.1016/j.mbs.2020.108320.

18. Hokkaido Government. Situation of COVID-19 in Hokkaido Prefecture [in Japanese]; accessed 2 April 2020. Available from: `http://www.pref.hokkaido.lg.jp/hf/kth/kak/hasseijoukyou.htm`.

19. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance. 2020;25(10). doi:10.2807/1560-7917.ES.2020.25.10.2000180.

20. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological assessment of hospitalized patients with COVID-2019. Nature. 2020;doi:10.1038/s41586-020-2196-x.

21. Eurosurveillance Editorial Team and others. Updated rapid risk assessment from ECDC on the novel coronavirus disease 2019 (COVID-19) pandemic: increased transmission in the EU/EEA and the UK. Eurosurveillance. 2020;25(10). doi:10.2807/1560-7917.ES.2020.25.10.2003121.

22. Pan J, Yao Y, Liu Z, Li M, Wang Y, Dong W, et al. Effectiveness of control strategies for Coronavirus Disease 2019: a SEIR dynamic modeling study. medRxiv. 2020;doi:10.1101/2020.02.19.20025387.

23. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). Science. 2020;doi:10.1126/science.abb3221.

24. Ministry of Health, Labour, and Welfare, Japan. Press release on March 15th, 2020 [in Japanese]; accessed 9 July 2020. Available from: `https://www.mhlw.go.jp/content/10900000/000608893.pdf`.

25. Nishiura H, Kobayashi T, Suzuki A, Jung SM, Hayashi K, Kinoshita R, et al. Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). International Journal of Infectious Diseases. 2020;doi:10.1016/j.ijid.2020.03.020.

26. Akhmetzhanov AR, Mizumoto K, Jung Sm, Linton NM, Omori R, Nishiura H. Estimation of the actual incidence of coronavirus disease (COVID-19) in emergent hotspots: The example of Hokkaido, Japan during February-March 2020. medRxiv. 2020;.

27. Hokkaido Government. Press release on March 21th, 2020 [in Japanese]; accessed 4 April 2020. Available from: `http://www.pref.hokkaido.lg.jp/hf/kth/kak/kisyakaiken03212sokuhou.pdf`.

28. The Japan Times. Hokkaido set to lift coronavirus state of emergency; accessed 2 April 2020. Available from: `https://www.japantimes.co.jp/news/2020/03/19/national/hokkaido-set-to-lift-coronavirus-emergency/`.

29. Kupferschmidt K. Why do some COVID-19 patients infect many others, whereas most don't spread the virus at all. Science. accessed 11 July 2020;.

30. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438(7066):355–359. doi:10.1038/nature04153.

31. Kucharski AJ, Althaus CL. The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. Eurosurveillance. 2015;20(25). doi:10.2807/1560-7917.ES2015.20.25.21167.

32. Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. Physical Review Letters. 2001;86(14):3200–3203. doi:10.1103/PhysRevLett.86.3200.

33. Moreno Y, Pastor-Satorras R, Vespignani A. Epidemic outbreaks in complex heterogeneous networks. The European Physical Journal B-Condensed Matter and Complex Systems. 2002;26(4):521–529. doi:10.1140/epjb/e20020122.

34. Rothe C, Schunk M, Sothmann P, Bretzel G, Froeschl G, Wallrauch C, et al. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. New England Journal of Medicine. 2020;382(10):970–971. doi:10.1056/NEJMc2001468.