Title: Shrinkage in serial intervals across cluster transmission generations of COVID-19

Shi Zhao[1,2,*], Yu Zhao[3], Biao Tang[4,5], Daozhou Gao[6], Zihao Guo[1], Marc KC Chong[1,2], Salihu S Musa[7,8], Yongli Cai[9], Weiming Wang[9], Daihai He[7,*], and Maggie H Wang[1,2]

1 JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong, China

2 CUHK Shenzhen Research Institute, Shenzhen, China

3 School of Public Health and Management, Ningxia Medical University, Yinchuan, Ningxia, China

4 School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China

5 Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3, Canada

6 Department of Mathematics, Shanghai Normal University, Shanghai, China

7 Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, China

8 Department of Mathematics, Kano University of Science and Technology, Wudil, Nigeria

9 School of Mathematics and Statistics, Huaiyin Normal University, Huaian, China

***** Correspondence to: zhaoshi.cmsa@gmail.com (SZ), or daihai.he@polyu.edu.hk (DH).

Email addresses of all authors

SZ: zhaoshi.cmsa@gmail.com; YZ: zhaoyuzy123@163.com; BT: btang66@yorku.ca; DG: dzgao@shnu.edu.cn; ZG: guozihao9602@163.com; MKCC: marc@cuhk.edu.hk; SSM: salihu-sabiu.musa@connect.polyu.hk; YC: yonglicai@hytc.edu.cn; WW: weimingwang2003@163.com; DH: daihai.he@polyu.edu.hk; MHW: maggiew@cuhk.edu.hk.

Abstract

The COVID-19 pandemic poses a serious threat to global health, and one of the key epidemiological factors that shape the transmission of COVID-19 is its serial interval (SI). Although SI is commonly considered following a probability distribution at a population scale, slight discrepancies in SI across different transmission generations are observed from the aggregated statistics in recent studies. To explore the change in SI across transmission generations, we develop a likelihood-based statistical inference framework to examine and quantify the change in SI. The COVID-19 contact tracing surveillance data in Hong Kong are used for exemplification. We find that the individual SI of COVID-19 is likely to shrink with a rate of 0.72 per generation and 95%CI: (0.54, 0.96) as the transmission generation increases. We speculate that the shrinkage in SI is an outcome of competition among multiple candidate infectors within a cluster of cases. The shrinkage in SI may speed up the transmission process, and thus the nonpharmaceutical interventive strategies are crucially important to mitigate the COVID-19 epidemic.

Keywords: COVID-19; serial interval; transmission generation; contact tracing; statistical modelling.

1 Introduction

The transmission dynamics of an infectious disease are largely determined by the pathogen's infectiousness and the course of the transmission [1-9]. The serial interval (SI), which is defined as the time interval between the symptoms onset dates of an infector and of the associated infectee [10-13], is widely used to measure the duration of the transmission generation. As the most efficient proxy of the generation time (GT) [14], SI is one of the crucial epidemiological parameters in shaping the transmission process as well as the growth patterns of an outbreak [7, 9, 15, 16].

As a contagious disease, the coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was firstly reported in 2019 [17-21], and rapidly spread to over 200 countries and territories, which poses a serious threat to global health. In response to the ongoing COVID-19 pandemic, the World Health Organization (WHO) declared a public health emergency of international concern on January 30, 2020 [22]. As of September 25, 2020, there have been over 31 million confirmed COVID-19 cases worldwide with over 0.9 million associated deaths [23].

To date, the transmission process of COVID-19 has been characterized and reconstructed both empirically and theoretically [6, 8, 18, 24-29]. In a number of existing literature, SI is commonly considered following a universal distribution at the population (or herd) scale for many well-known respiratory infectious diseases [13, 30-32], which also occurs for COVID-19 [6, 18, 33, 34]. Recently, two studies, both of which are based on the aggregated SI observations, reported that SI appears with slight discrepancies across different transmission generations [35, 36]. Inspiring by their findings, we suspect there may exist a solid difference in the infector's mean SI in consecutive generations in a transmission chain.

In this study, we develop a statistical framework to explore the change in SI across transmission generations. For exemplification, we quantify the change in SI by using the COVID-19 contact tracing surveillance data in Hong Kong. We explore the mechanism that drives the change in SI, and we also demonstrate its effects on shaping the transmission of COVID-19.

2 Methods

2.1 Conceptualization and statistical parameterization

We denote the SI of an infected individual, i.e., infector, by τ that follows a probability density function (PDF) $h(\tau)$ with mean μ and standard deviation (SD) σ. A transmission chain is composed by two consecutive transmission pairs, in which the infectee in the former transmission pair acts as the infector in the latter transmission pair, see Fig 1. For convenience, we name the SI in the former transmission pair by former SI and denoted by $\tau^{(F)}$, and the SI in the latter transmission pair by latter SI and denoted by $\tau^{(L)}$. Here, we note that the superscript, i.e., '(F)' or '(L)', is used merely as a label instead of as a power.

We explore the changing patterns in SI across transmission generations. In the same transmission chain, an intuitive statistical relation between $\tau^{(F)}$ and $\tau^{(L)}$ in Eqn (1),

$$\mathbf{E}[\tau^{(L)}] = \lambda \cdot \mathbf{E}[\tau^{(F)}], \quad (1)$$

is considered, where **E**[·] denotes the expectation function. The parameter λ is the change ratio between the means of two consecutive SIs, which is a positive constant to be determined. Straightforwardly, there exist iterative changes in mean SI across transmission generations, if $\lambda \neq 1$, while the mean SI may be a constant, if $\lambda = 1$. Hence, the relation in Eqn (1) can be examined by checking whether $\lambda = 1$ holds under the null hypothesis.

2.2 Likelihood-based inference framework

With the PDF $h(\tau)$ for the individual SI, the (baseline) likelihood framework, denoted by L_0, can be formulated in Eqn (2). That is

$$L_0(\lambda)=\prod_i \left[h^{(L)}\left(\tau_i^{(L)}\middle|\lambda,\tau_i^{(F)}\right)\cdot h^{(F)}\left(\tau_i^{(F)}\middle|\lambda,\tau_i^{(L)}\right)\right], \tag{2}$$

where the subscript i denotes the i-th transmission chain. For $h^{(L)}(\cdot|\lambda, \tau^{(F)})$, the mean of $h^{(L)}$ is given as $\mu^{(L)} = \tau^{(F)}\cdot\lambda$ according to the relation in Eqn (1). By contrast, for $h^{(F)}(\cdot|\lambda, \tau^{(L)})$, the mean of $h^{(F)}$ is given as $\mu^{(F)} = \tau^{(L)}/\lambda$. The SD of $h(\cdot)$, i.e., σ, is modelled as a function of μ. Due to the lack of information about the dispersion of the individual SI, we consider three scenarios of σ that cover a wide range of the possible situations. They include

- scenario (**I**), a large SD: $\sigma = |\mu|$, which refers to the scale of the coefficient of variation (CV) estimated in previous studies [6, 8, 24, 26, 29, 33, 37-40] and considered as an upper bound of SD;
- scenario (**II**), a moderate SD: $\sigma^2 = |\mu|$, which is assumed having a Poisson-like feature; and
- scenario (**III**), a small SD: $\sigma = 1$, which is assumed and considered as a lower bound of SD.

With the mean and SD, the function $h(\cdot)$ can be formulated by some widely adopted PDFs. We consider three different PDFs. They are

- Normal distribution as a representative of symmetric distributions defined on all real numbers [8, 35, 37-39, 41, 42];
- Gumbel distribution as a representative of asymmetric distributions defined on all real numbers [8, 37]; and
- Gamma distribution as a representative of asymmetric distributions defined on positive numbers [6, 8, 13, 18, 28-30, 33-35, 37, 38, 40, 43-45].

We select the scenario of SD and distribution of $h(\cdot)$ according to the fitting performance in terms of the Akaike information criterion with a correction for small sample sizes (AICc).

In addition, as pointed out in [33], the baseline likelihood in Eqn (2) might lead to an underestimation of SI due to the interval-censoring issue. Hence, according to the truncation scheme previously developed in [40], which accounts for the effects of each infector's isolation, we adjust for the right truncation bias by an improved likelihood function, L, in Eqn (3). We have

$$L(\lambda)=\prod_i \left[\frac{h^{(L)}\left(\tau_i^{(L)}\middle|\lambda,\tau_i^{(F)}\right)}{H^{(L)}\left(d_i^{(L)}\middle|\lambda,\tau_i^{(F)}\right)}\cdot\frac{h^{(F)}\left(\tau_i^{(F)}\middle|\lambda,\tau_i^{(L)}\right)}{H^{(F)}\left(d_i^{(F)}\middle|\lambda,\tau_i^{(L)}\right)}\right], \tag{3}$$

where $H(\cdot)$ is the cumulative distribution function (CDF) of $h(\cdot)$, and the letter d denotes the duration from the onset date of an infector to the person's isolation date. All other notations are the same as those in Eqn (2).

The parameter λ is estimated under both truncated and non-truncated schemes by using the maximum likelihood estimation (MLE). The AICc is employed for model selection. The 95% confidence interval (95%CI) is calculated by using the profile likelihood estimation framework with the cutoff threshold of a Chi-square quantile [46-50].

All analyses are conducted in the **R** statistical software (version 3.5.1), and no specific package is used.

2.3 COVID-19 surveillance data in Hong Kong

The COVID-19 surveillance data are originally released by the Centre for Health Protection (CHP) of Hong Kong [51], and used in [24] previously. According to the data description in [24], a total of 1038 laboratory-confirmed SARS-CoV-2 infections as of May 7, 2020, were initially screened. In Hong Kong, each contact of a confirmed COVID-19 case, defined as who has prolonged face-to-face interaction with a case, is traced and mandatorily quarantined for 14 days, regardless of symptom appearance. Then, each transmission pair, i.e., the 'infector-and-infectee' pair, can be reconstructed from the contact tracing records. A total of 169 transmission pairs including 27 asymptomatic transmission pairs for either infector or infectee, which are directly collected via https://github.com/dcadam/covid-19-sse/blob/master/data/transmission_pairs.csv, are identified for further screening.

In this study, we focus on the (169 – 27 =) 142 symptomatic transmission pairs in Hong Kong. We identify the infectee who acts as an infector in other transmission pairs, i.e., the 'secondary case' in Fig 1, by matching all combinations of the 142 transmission pairs. We reconstructed the transmission chain with 3 generations including primary case, secondary case, and tertiary case, which is illustrated as the 'secondary case' in Fig 1. A total of 21 transmission chains are extracted, and presented in Fig 2.

Since the isolation period of each infector is unavailable, we consider the case confirmation date as a proxy of the isolation starting time with the presumption that the isolation starts immediately after confirmation. Hence, the change ratio of SI, λ, can be estimated from these transmission chain data in Hong Kong by using the analytical framework in Section 2.2.

2.4 Sensitivity analysis

To evaluate the estimating sensitivity, an alternative formulation, similar to the relationship in Eqn (1), is adopted to repeat the estimation with the dataset from Hong Kong. The alternative relationship between the former and latter SIs is formulated in Eqn (4).

$$\mathbf{E}[\tau^{(\mathrm{L})}] = \lambda \cdot [\mathbf{E}[\tau^{(\mathrm{F})}] - T_{\mathrm{c}}] + T_{\mathrm{c}}, \tag{4}$$

where the term T_{c} (≥ 0) indicates the lower bound of the SI as generation increases. Other terms have the same meanings as those in Eqn (1). Straightforwardly, Eqn (1) and Eqn (4) will be equivalent, if $T_{\mathrm{c}} = 0$. Thus, the intuition of Eqn (4) is of the same fashion as that of the Eqn (1).

We estimate both T_c and λ simultaneously with the likelihood profiles and estimation procedures in Section 2.2. The model selection is conducted referring to the lowest AICc. We check the consistency of the λ estimates, and whether T_c is significantly larger than 0.

2.5 Exploring the mechanisms behind change in SI

In this section, we develop statistical models to explore two possible, but not verified, mechanisms behind the change in SI, their effects in shaping the transmission process, and their reasonability.

2.5.1 Exploration #1: changes in latent period and infectious period

In exploration #1, we consider a hypothetical scenario that the change in SI is an intrinsic feature of the pathogen, which is due to change in latent period and infectious period across cluster generations. Then, according to the classic 'susceptible-exposed-infectious-removed' (SEIR) framework, where exponential distributions are assumed for most of the epidemiological parameters [52-56], we have

$$X^{(k)} + Y^{(k)} = \mathbf{E}[\tau^{(k)}], \text{ and } X^{(k+1)} + Y^{(k+1)} = \mathbf{E}[\tau^{(k+1)}], \quad (4)$$

where X (unit: day) denotes the mean latent period, and Y (unit: day) denotes the infectious period. The superscript (k) is the label of transmission generation rather than a power. This relationship is derived in [54] theoretically, and adopted in [15, 52, 53, 55, 56]. When $\lambda < 1$, we assume $0 \le X^{(k+1)} \le X^{(k)}$, and $0 \le Y^{(k+1)} \le Y^{(k)}$ for Eqn (4). We define $\rho^{(k)} = \frac{Y^{(k)} - Y^{(k+1)}}{E\left[\tau^{(k)}\right] - E\left[\tau^{(k+1)}\right]} \times 100\,\%$ as the percentage of SI reduction due to the reduction in infectious period.

We explore the potential effects of the change in SI on the individual reproduction number, R, across transmission generations. Referring to the SEIR framework, the individual reproduction number can be modelled as the product of the mean effective contact rate and the mean infectious period, i.e., $R^{(k)} = \beta^{(k)} \cdot Y^{(k)}$ for the infector in the k-th generation in a transmission chain. Here, β (unit: per day) denotes the effective contact rate.

By fixing β as a constant, we explore the effects of the change in τ on R in the k-th transmission generation. To set up, we fix the mean SI of the infector, $\mathbf{E}[\tau^{(k=0)}]$, at 7.5 days referring to the estimates from the earliest COVID-19 data [18], and the mean latent period, $X^{(k=0)}$, at 3.3 days [45, 57] for the initial, i.e., 0-th, generation. Thus, the mean infectious period, $Y^{(k=0)}$, is derived at (7.5 − 3.3=) 4.3 days by using Eqn (4), which is in line with the results in literatures [55, 57, 58]. We further fix the initial individual reproduction number, $R^{(k=0)}$, at 2.2, which is generally consistent with previous estimates [3, 6, 8, 18, 19, 37, 40, 52, 55, 59-62]. Thus, the relationship among k, ρ, and R can be solved numerically.

2.5.2 Exploration #2: competition among multiple candidate infectors

In exploration #2, we consider a statistical mechanism that the shrinkage in SI may be an outcome of a competition among multiple candidate infectors. The SI is recorded pairwisely as the duration between onset dates of an infectee and the infector who triggers the infection. In a cluster of cases, contacts are likely to occur in most pairs of infected and susceptible individuals simultaneously. The candidate infector is defined as those cases who contribute to the exposure of an

infectee but may or may not trigger the infection eventually. We speculate the competitions among multiple candidate infectors may shorten the SI.

For the competition among a total of J candidate infectors for one infectee, the onset time, t, of the infectee who is triggered by the j-th candidate infector follows a PDF denoted by $g(t = t_j + \tau_j)$. Here, t_j denotes onset date of the j-th candidate infector, and τ_j denotes the candidate SI, if occurs, between the j-th candidate infector and the infectee. The parameter t_j is observable from the surveillance data, and thus is considered as a constant. The parameter τ_j is modelled as independent and identically distributed (IID) random variable following the PDF $h(\tau)$ as defined in Section 2.1. Hence, the $g(t)$ appears a shifted version of $h(\tau)$ with a shift term of $-t_j$. The candidate infector who triggers the infectee at the earliest is recognized as the infector. Thus, the observed SI of the infectee is τ_j that associates with the smallest $(t_j + \tau_j)$ for all indexes j.

We simulate this candidate infector competition framework stochastically. To set up, we consider a cluster starting with one seed case whose onset date is day (or time) 0. The PDF $h(\tau)$ is modelled as a Gamma distribution with mean 5.5 and standard deviation (SD) 3.3 days, which is in line with many existing estimates [29, 43-45]. With h, the PDF g can also be determined by shifting. For the reproducibility, we restrict the number of offsprings generated by each infector following a Poisson distribution with rate parameter fixed at 2.2, which is consistent to the predefined value of reproduction number (R) in Section 2.5.1. The number of offsprings can also be directly assigned manually. Specially, the number of offsprings from the initial seed case, namely number of primary offsprings, is a criterion to identify the superspreading events, and thus is importance to explore its effect in shaping SI across transmission generation. For simplicity, we neglect the isolation in the simulation framework.

During the model simulation, we record the cluster size in terms of the cumulative number of cases, onset dates of each case, infector of each infectee (except the initial seed case), SI, generation of cases, and number of offsprings for each infector. The generation of cases is traced by the transmission chain linked to the initial seed case, and we defined the generation of initial seed case as generation 0. For convenience, the transmission generation between a case in generation 0 and another case in generation 1 as the first transmission generation, and thus the index of transmission generation can be ranked subsequently.

For each simulation, we extract the SIs from first and second transmission generations, and treat these consecutive SIs as pairs of former and latter SIs that is illustrated in Fig 1. We generate 30 pairs of former and latter SIs, and conduct the estimation of λ using the framework in Eqn (2). We explore the effects of cluster size, number of primary offsprings and generation numbers in changing the scale of SI.

3 Results and discussion

For the 21 identified COVID-19 transmission chains in Hong Kong, the pairs of former and latter SIs are presented in Fig 2. We report the descriptive statistics as follows. For the former SI, we report a mean of 5.4 days, median of 6.0 days, interquartile range (IQR) between 3.0 and 7.0 days, 95% centile from 1.5 to 10.5 days, 95% percentile of 8.0 days, and a range from 1.0 to 13.0 days. For the latter SI, we report a mean of 4.8 days, median of 4.0 days, IQR between 3.0 and 7.0 days, 95% centile from 1.0 to 9.5 days, 95% percentile of 9.0 days, and a range from 1.0 to 10.0 days. We

observe that the mean (and median) SI decreases when generation increases, and this finding was reported previously in [35, 36]. With the sample means, we calculate the ratio of latter SI over former SI at (4.8 / 5.4 =) 0.89, which is roughly the same scale as 0.73 in [35] and 0.94 or 0.75 in [36]. Empirically, the pairwise difference of latter SI minus former SI has a mean of −0.7 days, median of 0.0 day, and IQR between −3.0 and 2.0 days. The pairwise ratio of latter SI over former SI has a mean of 1.1, median of 1.0, and IQR between 0.5 and 1.5. By using the nonparametric bootstrapping approach, the crude change ratio of SI across generations is calculated at 1.00 with 95%CI: (0.57, 1.43).

Considering the theorical probability profile of individual SI in Eqn (2), we estimate the λ at 0.77 and 95%CI: (0.51, 1.16) selected with the lowest AICc among all non-truncated scenarios, see Table 1. For all scenarios in Table 1, we find that the Gamma distribution with $\sigma^2 = |\mu|$ outperformed against other scenarios in terms of the lowest AICc. As such, we estimate the λ at 0.72 and 95%CI: (0.54, 0.96), which are considered as the main results. Besides the fitting performance, we also consider the biological feasibility of probability profile in governing the real-world observations of the SI of COVID-19. Referring to the previous literatures [8, 24, 29, 37-39, 44, 45], the SI of COVID-19 might be negative, i.e., $\tau < 0$. Although the Gamma distribution outperforms, the negative SI observations cannot be governed by a Gamma-distributed $h(\cdot)$. In this case, the scenario with the second lowest AICc is considered as another main results. As such, we estimate the λ at 0.74 and 95%CI: (0.61, 0.91) with a Gumbel distribution, which is also highlighted in Table 1. The best fitting performance from Gamma distribution is probably because all our SI observations appear positive, see Fig 2. We remark that with negative SI observations, Gumbel distribution is likely to yield a better fitting performance than Gamma distribution.

Consistently, the estimates of λ using Gamma and Gumbel distributions are almost the same, and significantly less than 1. Thus, the individual SI is likely to shrink when the transmission generation increases with rate at 0.72 per generation, which, to the best of our knowledge, is the first study quantifying this relation. According to our truncated likelihood framework in Eqn (3), the estimated shrinkage can be understood in regard to the intrinsic SI [33, 40], of which the "*distribution depends only on the average infectiousness of an individual*" as defined in [15]. Hence, the shrinkage in intrinsic SI is interpreted as an across-generation feature, and is unlikely due to the effects of nonpharmaceutical interventions, which may shorten the realized SI as pointed out in [37, 63-65]. Regardless of the number of direct offspring in each generation, the shrinkage in SI implies that the transmission is likely to occur more rapidly since the exposure of each infector. Then, the infectee is more likely exposed before the symptom onset of the infector in the late generations, comparing to the situation in the early generations. In other words, pre-symptomatic transmission may occur more frequently in the late generations. In addition, we find both main estimates of λ are under the scenario (**II**) of individual SI's SD (σ). Since the SIs from a population may have an SD as in scenario (**I**), this finding indicates that the SI of a population is more dispersive than the individual SI.

For the sensitivity analysis, the relationship in Eqn (4) is examined. We find that, consistent with the main results, the Gamma-distributed $h(\cdot)$ with scenario (**II**) of σ and likelihood truncation outperforms among other scenarios, see Fig 3. The SI lower bound, T_c, is estimated at 0.0 exactly,

which means Eqn (4) becomes equivalent to Eqn (1) and implies relationship in Eqn (1) holds robustly.

We explore the impacts of the shrinkage in SI in shaping the individual reproduction number modelled as exploration #1 in Section 2.5.1. We find that the R decreases when the transmission generations increase, see Fig 4. With a higher percentage of the reduction in SI due to the reduction in infectious period (ρ), the R may decrease more rapidly. When ρ is fixed at 0.5, 50% of the reduction in SI is from the reduction in infectious period (Y), and the remaining 50% is from the reduction in latent period (X). With the conditions for COVID-19 in Section 2.5.1 and λ at 0.72 per generation, we calculate the R of 1.1 in the second generation and of 0.9 in the third generation. The total number of offsprings seeded by an index case, commonly known as patient zero (in generation zero), will be 15.5 on average, if no control measure is implemented, which is 7.7-fold of the initial reproduction number as fixed at 2.2. Therefore, both timely and effective case's isolation and close contact's quarantine are crucial to mitigate the COVID-19 epidemics.

As a geometric sequence, if the absolute value of the common ratio, i.e., λ, is less than one, the sequence defined in Eqn (1) will converge. Thus, the mean SI will decrease and approach 0 theoretically, when the number of transmission generations becomes sufficiently large. However, under exploration #1, there occurs a discrepancy as follows.

> When the SI decreases, the individual R of each infector will also decrease, which leads to an outcome that the transmission of COVID-19 may vanish after a number of generations. However, as a matter of fact, the pandemic of COVID-19 continuous and to date, still maintains growing patterns in the epidemic curve.

We notice this discrepancy between the theoretical outcome from exploration #1 and the real-world observation.

For the suspected candidate infector competition mechanism proposed as exploration #2 in Section 2.5.2, we find the shrinkage in SI is likely occur when the cluster size increases, see Fig 5, and when the number of offsprings increases, see Fig 6. Under the mechanism in exploration #2, the mean individual reproduction number holds as a constant. In other words, the outbreak maintains with substantial offspring cases in each transmission generation, and thus the discrepancy occurring under exploration #1 vanishes. Therefore, we consider exploration #2 as the main discussion, which may be more reasonable than exploration #1.

We observe the SI shrinks as generation increases, and approaches a boundary level in the late generations, see Fig 7. As such, we argue that the alternative relation in Eqn (4) may be more biologically reasonable, even though the simpler formulation in Eqn (1) slightly outperforms. We speculate the outperformance of Eqn (1) is possibly because most transmission chains (16 out of 21) are the chains of cases from 'zero-first-second' generations in each cluster of COVID-19 cases. This character of our COVID-19 dataset makes the simple geometric relation in Eqn (1) an optimal fit to the observations from early generations. In other words, if more SI observations from late generations would be included, Eqn (4) may replace Eqn (1) as the optimal relationship. To verify, we repeat the estimation in Section 2.4 by solely using the (21 – 16 =) 5 transmission chains that are from late generations, i.e., secondary, tertiary, or quaternary. In this case, we estimate the T_c at 1.4 days (data not shown), which indicates Eqn (4) appears more feasible than Eqn (1). With a strictly

positive T_c as the lower bound of mean SI, the individual R may maintain at or over a certain level such that the COVID-19 outbreak continues. Hence, we remark that the data with more generations observed from each transmission chain will probably improve the estimation of the change in SI across generations.

Within a cluster of COVID-19 cases, the shrinkage in SI may speed up the transmission process, and thus result in increase in growth rate of the epidemic curve. The nonpharmaceutical interventive strategies, which can cut off the transmission chain, e.g., isolation, quarantine, social distancing, and personal protective equipment (PPE), are thus crucially important to mitigate the cluster size and flattening the epidemic curve. The statistical mechanism in exploration #2 is applicable to study the transmission dynamics of other infectious diseases. Future studies on verifying the exploration #2, or on exploring other clinical or biological mechanisms that affects the individual SI across transmission generations are desired.

4 Conclusions

The individual SI of COVID-19 is likely to shrink as the transmission generation increases. We speculate that the shrinkage in SI is an outcome of competition among multiple candidate infectors within a cluster of cases. The shrinkage in SI may speed up the transmission process, and thus the nonpharmaceutical interventive strategies are crucially important to mitigate the epidemic.

Declarations

Ethics approval, consent to participate, and consent for publication

All data used in this work are publicly available, and thus neither ethical approval nor consent is applicable.

Availability of materials

The COVID-19 surveillance data are collected via https://github.com/dcadam/covid-19-sse/blob/master/data/transmission_pairs.csv, which are originally released by the Centre for Health Protection (CHP) of Hong Kong [51] and previously used in [24].

Funding

DH was supported by General Research Fund (grant number: 15205119) of the Research Grants Council (RGC) of Hong Kong, China, and Alibaba (China) Co. Ltd. Collaborative Research project. YC and WW were supported by the National Natural Science Foundation of China (grant numbers 61672013, and 12071173), and the Huaian Key Laboratory for Infectious Diseases Control and Prevention (HAP201704).

Acknowledgements

None.

Disclaimer

The funding agencies had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Conflict of interests

DH received funding from Alibaba (China) Co. Ltd. Collaborative Research project. MHW is a shareholder of Beth Bioinformatics Co., Ltd. Other authors declared no competing interests.

Authors' contributions

SZ conceived the study, carried out the analysis, and drafted the first manuscript. SZ and DH discussed the results. All authors critically read and revised the manuscript, and gave final approval for publication.

References

1. Tuite AR, Fisman DN: **Reporting, Epidemic Growth, and Reproduction Numbers for the 2019 Novel Coronavirus (2019-nCoV) Epidemic**. *Annals of internal medicine* 2020, **172**(8):567-568.
2. Zhao S, Cao P, Gao D, Zhuang Z, Cai Y, Ran J, Chong MKC, Wang K, Lou Y, Wang W *et al*: **Serial interval in determining the estimation of reproduction number of the novel coronavirus disease (COVID-19) during the early outbreak**. *J Travel Med* 2020, **27**(3).
3. Riou J, Althaus CL: **Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020**. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2020, **25**(4).
4. Kutter JS, Spronken MI, Fraaij PL, Fouchier RA, Herfst S: **Transmission routes of respiratory viruses among humans**. *Curr Opin Virol* 2018, **28**:142-151.
5. Zhao S: **To avoid the noncausal association between environmental factor and COVID-19 when using aggregated data: Simulation-based counterexamples for demonstration**. *Sci Total Environ* 2020:141590.
6. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, Lau YC, Wong JY, Guan Y, Tan X: **Temporal dynamics in viral shedding and transmissibility of COVID-19**. *Nat Med* 2020:1-4.
7. Wallinga J, Lipsitch M: **How generation intervals shape the relationship between growth rates and reproductive numbers**. *Proceedings of the Royal Society B: Biological Sciences* 2007, **274**(1609):599-604.
8. Xu XK, Liu XF, Wu Y, Ali ST, Du Z, Bosetti P, Lau EHY, Cowling BJ, Wang L: **Reconstruction of Transmission Pairs for novel Coronavirus Disease 2019 (COVID-19) in mainland China: Estimation of Super-spreading Events, Serial Interval, and Hazard of Infection**. *Clin Infect Dis* 2020.
9. Yan P: **Separate roles of the latent and infectious periods in shaping the relation between the basic reproduction number and the intrinsic growth rate of infectious disease outbreaks**. *Journal of Theoretical Biology* 2008, **251**(2):238-252.
10. Fine PE: **The interval between successive cases of an infectious disease**. *American journal of epidemiology* 2003, **158**(11):1039-1047.
11. White LF, Wallinga J, Finelli L, Reed C, Riley S, Lipsitch M, Pagano M: **Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA**. *Influenza Other Respir Viruses* 2009, **3**(6):267-276.
12. Milwid R, Steriu A, Arino J, Heffernan J, Hyder A, Schanzer D, Gardner E, Haworth-Brockman M, Isfeld-Kiely H, Langley JM *et al*: **Toward Standardizing a Lexicon of Infectious Disease Modeling Terms**. *Front Public Health* 2016, **4**(213):213.
13. Vink MA, Bootsma MC, Wallinga J: **Serial intervals of respiratory infectious diseases: a systematic review and analysis**. *American journal of epidemiology* 2014, **180**(9):865-875.
14. Wallinga J, Teunis P: **Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures**. *American journal of epidemiology* 2004, **160**(6):509-516.
15. Champredon D, Dushoff J: **Intrinsic and realized generation intervals in infectious-disease transmission**. *Proceedings Biological sciences* 2015, **282**(1821):20152026.
16. Kenah E, Lipsitch M, Robins JM: **Generation interval contraction and epidemic data analysis**. *Mathematical biosciences* 2008, **213**(1):71-79.
17. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X *et al*: **Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China**. *Lancet (London, England)* 2020, **395**(10223):497-506.
18. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY *et al*: **Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia**. *N Engl J Med* 2020, **382**(13):1199-1207.
19. Zhao S, Musa SS, Lin Q, Ran J, Yang G, Wang W, Lou Y, Yang L, Gao D, He D *et al*: **Estimating the Unreported Number of Novel Coronavirus (2019-nCoV) Cases in China in the First Half of January 2020: A Data-Driven Modelling Analysis of the Early Outbreak**. *J Clin Med* 2020, **9**(2):388.

20. Leung K, Wu JT, Liu D, Leung GM: **First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment**. *Lancet (London, England)* 2020, **395**(10233):1382-1393.
21. Parry J: **China coronavirus: cases surge as official admits human to human transmission**. *BMJ (Clinical research ed)* 2020, **368**:m236.
22. **Statement on the second meeting of the International Health Regulations Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV), World Health Organization (WHO).** [https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov)]
23. **Novel Coronavirus (2019-nCoV) situation reports, released by the World Health Organization (WHO).** [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports]
24. Adam DC, Wu P, Wong JY, Lau EHY, Tsang TK, Cauchemez S, Leung GM, Cowling BJ: **Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong**. *Nat Med* 2020.
25. Luo L, Liu D, Liao X, Wu X, Jing Q, Zheng J, Liu F, Yang S, Bi H, Li Z *et al*: **Contact Settings and Risk for Transmission in 3410 Close Contacts of Patients With COVID-19 in Guangzhou, China : A Prospective Cohort Study**. *Annals of internal medicine* 2020.
26. Kwok KO, Wong VWY, Wei WI, Wong SYS, Tang JW: **Epidemiological characteristics of the first 53 laboratory-confirmed cases of COVID-19 epidemic in Hong Kong, 13 February 2020**. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2020, **25**(16):2000155.
27. Wu J, Huang Y, Tu C, Bi C, Chen Z, Luo L, Huang M, Chen M, Tan C, Wang Z *et al*: **Household Transmission of SARS-CoV-2, Zhuhai, China, 2020**. *Clin Infect Dis* 2020.
28. Ren X, Li Y, Yang X, Li Z, Cui J, Zhu A, Zhao H, Yu J, Nie T, Ren M *et al*: **Evidence for pre-symptomatic transmission of coronavirus disease 2019 (COVID-19) in China**. *Influenza Other Respir Viruses* 2020.
29. Tindale LC, Stockdale JE, Coombe M, Garlock ES, Lau WYV, Saraswat M, Zhang L, Chen D, Wallinga J, Colijn C: **Evidence for transmission of COVID-19 prior to symptom onset**. *Elife* 2020, **9**:e57149.
30. Cowling BJ, Fang VJ, Riley S, Malik Peiris JS, Leung GM: **Estimation of the serial interval of influenza**. *Epidemiology (Cambridge, Mass)* 2009, **20**(3):344-347.
31. Leung GM, Hedley AJ, Ho L-M, Chau P, Wong IOL, Thach TQ, Ghani AC, Donnelly CA, Fraser C, Riley S: **The epidemiology of severe acute respiratory syndrome in the 2003 Hong Kong epidemic: an analysis of all 1755 patients**. *Annals of internal medicine* 2004, **141**(9):662-673.
32. Assiri A, McGeer A, Perl TM, Price CS, Al Rabeeah AA, Cummings DA, Alabdullatif ZN, Assad M, Almulhim A, Makhdoom H *et al*: **Hospital outbreak of Middle East respiratory syndrome coronavirus**. *N Engl J Med* 2013, **369**(5):407-416.
33. Nishiura H, Linton NM, Akhmetzhanov AR: **Serial interval of novel coronavirus (COVID-19) infections**. *Int J Infect Dis* 2020, **93**:284-286.
34. Wang K, Zhao S, Liao Y, Zhao T, Wang X, Zhang X, Jiao H, Li H, Yin Y, Wang MH *et al*: **Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020**. *Transbound Emerg Dis* 2020.
35. Ma S, Zhang J, Zeng M, Yun Q, Guo W, Zheng Y, Zhao S, Wang MH, Yang Z: **Epidemiological parameters of coronavirus disease 2019: a pooled analysis of publicly reported individual data of 1155 cases from seven countries**. *J Med Internet Res* 2020:Accepted.
36. Li M, Liu K, Song Y, Wang M, Wu J: **Serial interval and generation interval for respectively the imported and local infectors estimated using reported contact-tracing data of COVID-19 in China**. *medRxiv* 2020.
37. Ali ST, Wang L, Lau EHY, Xu X-K, Du Z, Wu Y, Leung GM, Cowling BJ: **Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions**. *Science* 2020.
38. Du Z, Xu X, Wu Y, Wang L, Cowling BJ, Meyers LA: **Serial Interval of COVID-19 among Publicly Reported Confirmed Cases**. *Emerg Infect Dis* 2020, **26**(6):1341-1343.

39. You C, Deng Y, Hu W, Sun J, Lin Q, Zhou F, Pang CH, Zhang Y, Chen Z, Zhou XH: **Estimation of the time-varying reproduction number of COVID-19 outbreak in China**. *Int J Hyg Environ Health* 2020, **228**:113555.
40. Zhao S, Gao D, Zhuang Z, Chong MKC, Cai Y, Ran J, Cao P, Wang K, Lou Y, Wang W *et al*: **Estimating the Serial Interval of the Novel Coronavirus Disease (COVID-19): A Statistical Analysis Using the Public Data in Hong Kong From January 16 to February 15, 2020**. *Frontiers in Physics* 2020, **8**.
41. Yang L, Dai J, Zhao J, Wang Y, Deng P, Wang J: **Estimation of incubation period and serial interval of COVID-19: analysis of 178 cases and 131 transmission chains in Hubei province, China**. *Epidemiol Infect* 2020, **148**:e117.
42. Forsberg White L, Pagano M: **A likelihood−based method for real−time estimation of the serial interval and reproductive number of an epidemic**. *Stat Med* 2008, **27**(16):2999-3016.
43. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dorner L, Parker M, Bonsall D, Fraser C: **Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing**. *Science* 2020, **368**(6491):eabb6936.
44. Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, Hens N: **Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020**. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2020, **25**(17):2000257.
45. Zhao S: **Estimating the time interval between transmission generations when negative values occur in the serial interval data: using COVID-19 as an example**. *Mathematical Biosciences and Engineering* 2020, **17**(4):3512-3519.
46. Fan J, Huang T: **Profile likelihood inferences on semiparametric varying-coefficient partially linear models**. *Bernoulli* 2005, **11**(6):1031-1057.
47. Zhao S, Musa SS, Meng J, Qin J, He D: **The long-term changing dynamics of dengue infectivity in Guangdong, China, from 2008−2018: a modelling analysis**. *Trans R Soc Trop Med Hyg* 2020, **114**(1):62-71.
48. Lin Q, Chiu AP, Zhao S, He D: **Modeling the spread of Middle East respiratory syndrome coronavirus in Saudi Arabia**. *Stat Methods Med Res* 2018, **27**(7):1968-1978.
49. Cai Y, Zhao S, Niu Y, Peng Z, Wang K, He D, Wang W: **Modelling the effects of the contaminated environments on tuberculosis in Jiangsu, China**. *J Theor Biol* 2020, **508**:110453.
50. He D, Zhao S, Lin Q, Musa SS, Stone L: **New estimates of the Zika virus epidemic attack rate in Northeastern Brazil from 2015 to 2016: A modelling analysis based on Guillain-Barre Syndrome (GBS) surveillance data**. *PLoS Negl Trop Dis* 2020, **14**(4):e0007502.
51. **Summary of data and outbreak situation of the Severe Respiratory Disease associated with a Novel Infectious Agent, Centre for Health Protection, the government of Hong Kong.** [https://www.chp.gov.hk/en/features/102465.html]
52. Gatto M, Bertuzzo E, Mari L, Miccoli S, Carraro L, Casagrandi R, Rinaldo A: **Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures**. *Proceedings of the National Academy of Sciences of the United States of America* 2020, **117**(19):10484-10491.
53. Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, Gopalakrishna G, Chew SK, Tan CC, Samore MH *et al*: **Transmission dynamics and control of severe acute respiratory syndrome**. *Science* 2003, **300**(5627):1966-1970.
54. Svensson A: **A note on generation times in epidemic models**. *Mathematical biosciences* 2007, **208**(1):300-311.
55. Wu JT, Leung K, Leung GM: **Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study**. *Lancet (London, England)* 2020, **395**(10225):689-697.
56. Zhao S, Stone L, Gao D, Musa SS, Chong MKC, He D, Wang MH: **Imitation dynamics in the mitigation of the novel coronavirus disease (COVID-19) outbreak in Wuhan, China from 2019 to 2020**. *Ann Transl Med* 2020, **8**(7):448.

57. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J: **Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)**. *Science* 2020, **368**(6490):489-493.
58. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM, Centre for Mathematical Modelling of Infectious Diseases C-wg: **Early dynamics of transmission and control of COVID-19: a mathematical modelling study**. *The Lancet Infectious diseases* 2020, **20**(5):553-558.
59. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, Pastore YPA, Mu K, Rossi L, Sun K *et al*: **The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak**. *Science* 2020, **368**(6489):395-400.
60. Jung SM, Akhmetzhanov AR, Hayashi K, Linton NM, Yang Y, Yuan B, Kobayashi T, Kinoshita R, Nishiura H: **Real-Time Estimation of the Risk of Death from Novel Coronavirus (COVID-19) Infection: Inference Using Exported Cases**. *J Clin Med* 2020, **9**(2).
61. Musa SS, Zhao S, Wang MH, Habib AG, Mustapha UT, He D: **Estimation of exponential growth rate and basic reproduction number of the coronavirus disease 2019 (COVID-19) in Africa**. *Infect Dis Poverty* 2020, **9**(1):96.
62. Ran J, Zhao S, Han L, Liao G, Wang K, Wang MH, He D: **A re-analysis in exploring the association between temperature and COVID-19 transmissibility: an ecological study with 154 Chinese cities**. *Eur Respir J* 2020, **56**(2):2001253.
63. Nishiura H: **Time variations in the generation time of an infectious disease: implications for sampling to appropriately quantify transmission potential**. *Mathematical Biosciences & Engineering* 2010, **7**(4):851.
64. Zhao S, Cao P, Chong MKC, Gao D, Lou Y, Ran J, Wang K, Wang W, Yang L, He D *et al*: **COVID-19 and gender-specific difference: Analysis of public surveillance data in Hong Kong and Shenzhen, China, from January 10 to February 15, 2020**. *Infect Control Hosp Epidemiol* 2020, **41**(6):750-751.
65. Park SW, Champredon D, Dushoff J: **Inferring generation-interval distributions from contact-tracing data**. *Journal of the Royal Society Interface* 2020, **17**(167):20190719.

Figures

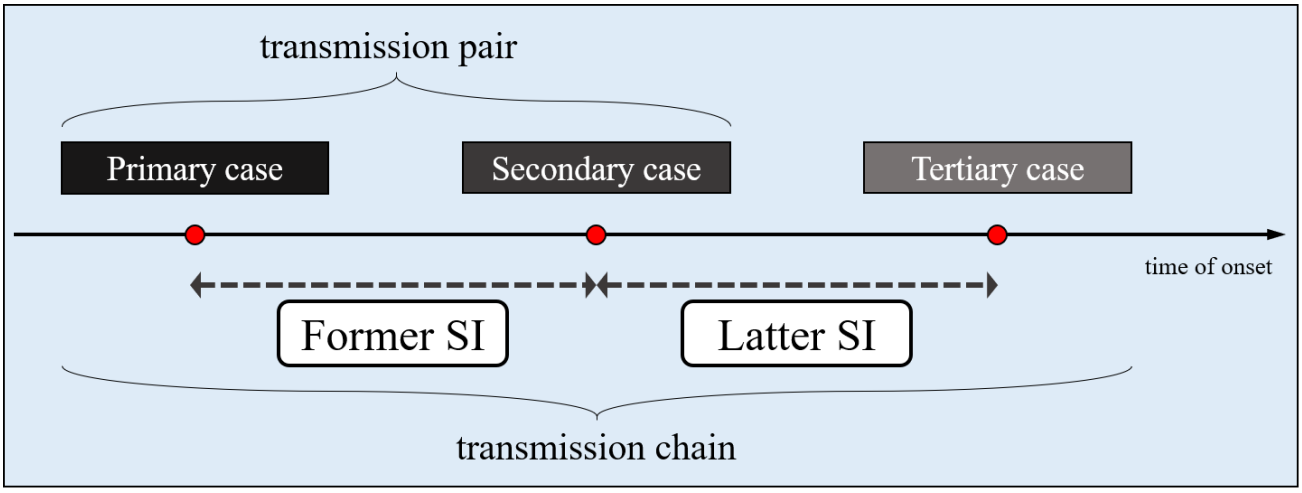


Figure 1.

The illustration diagram of the timeline of a typical transmission chain.

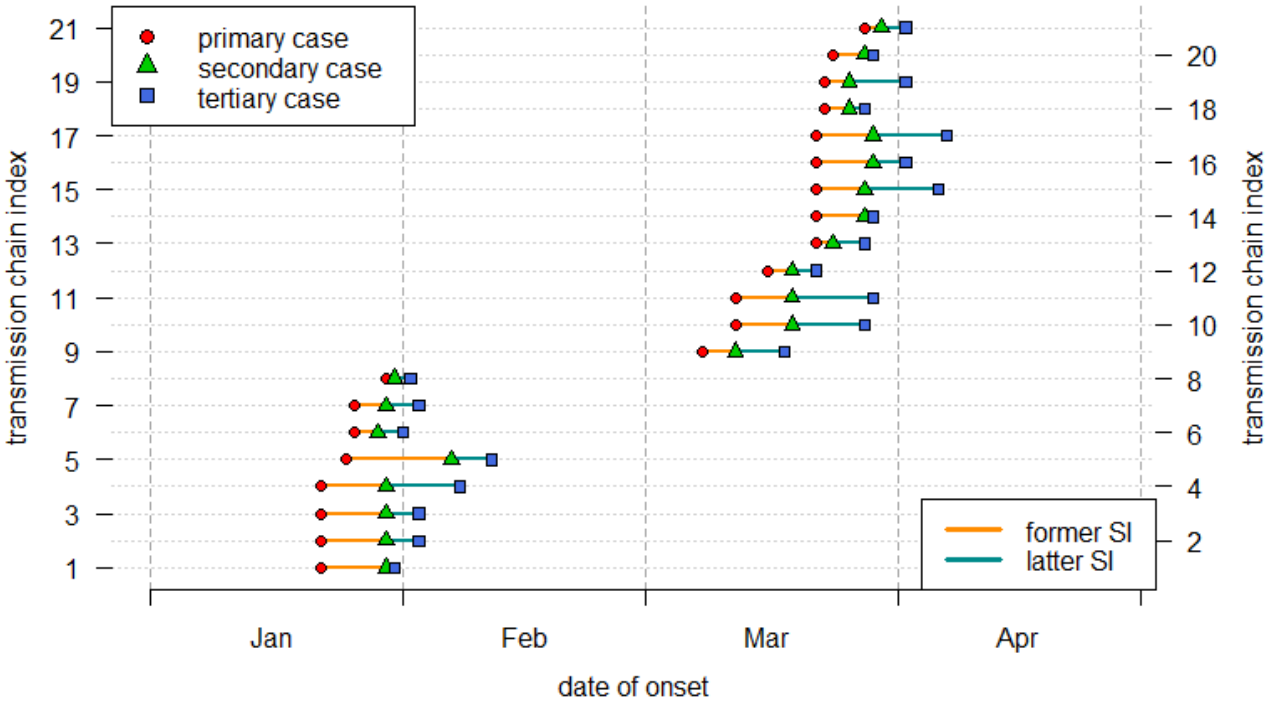
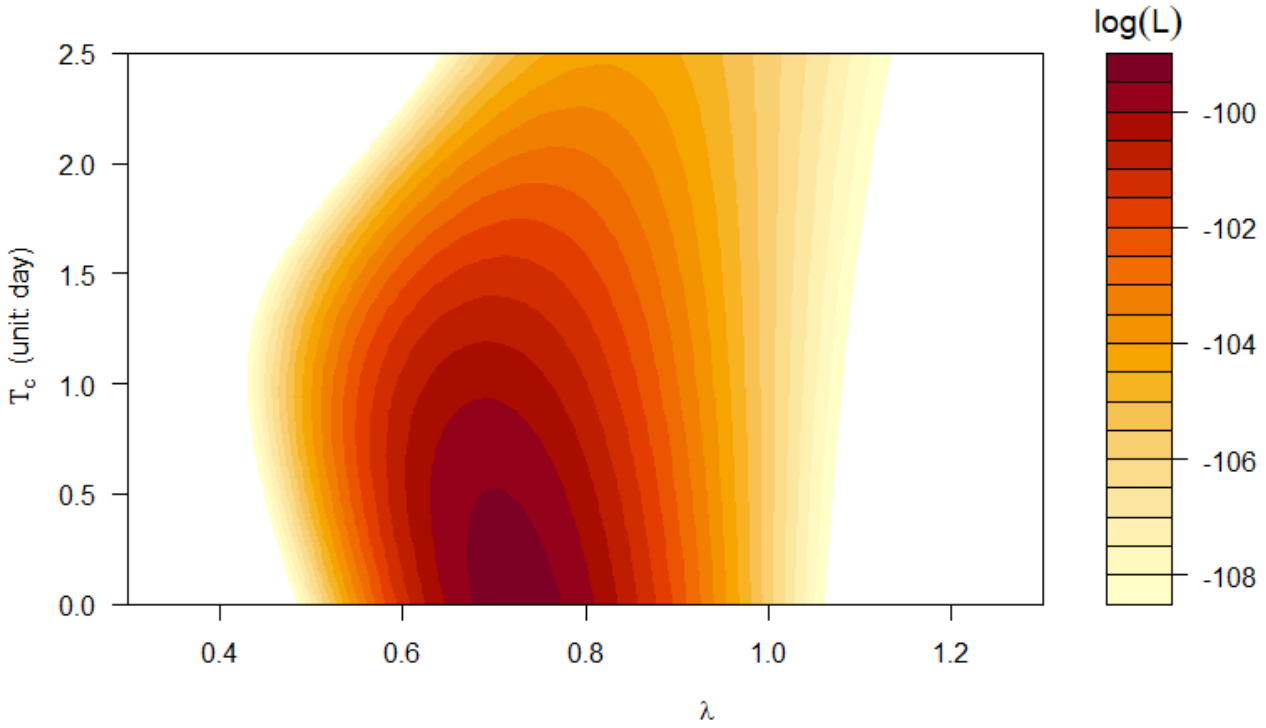


Figure 2.

The timeline of the transmission chains included in this study. The dot indicates the symptoms onset date of each case. The horizontal solid line represents the duration of each serial interval (SI). The transmission chains are indexed in the sequence of the onset dates of primary, secondary, and tertiary cases, which is merely for visualization purposed and will not affect the analytical procedures.



549

550 Figure 3.

551 The Gamma-distributed log-likelihood profile of T_c (unit: day) and λ, in Eqn (4), under scenario (**II**)
552 of σ, which has the best fitting performance in terms of the AICc = 202.8. The color scheme of the
553 log-likelihood values is shown in the right column.

554

555

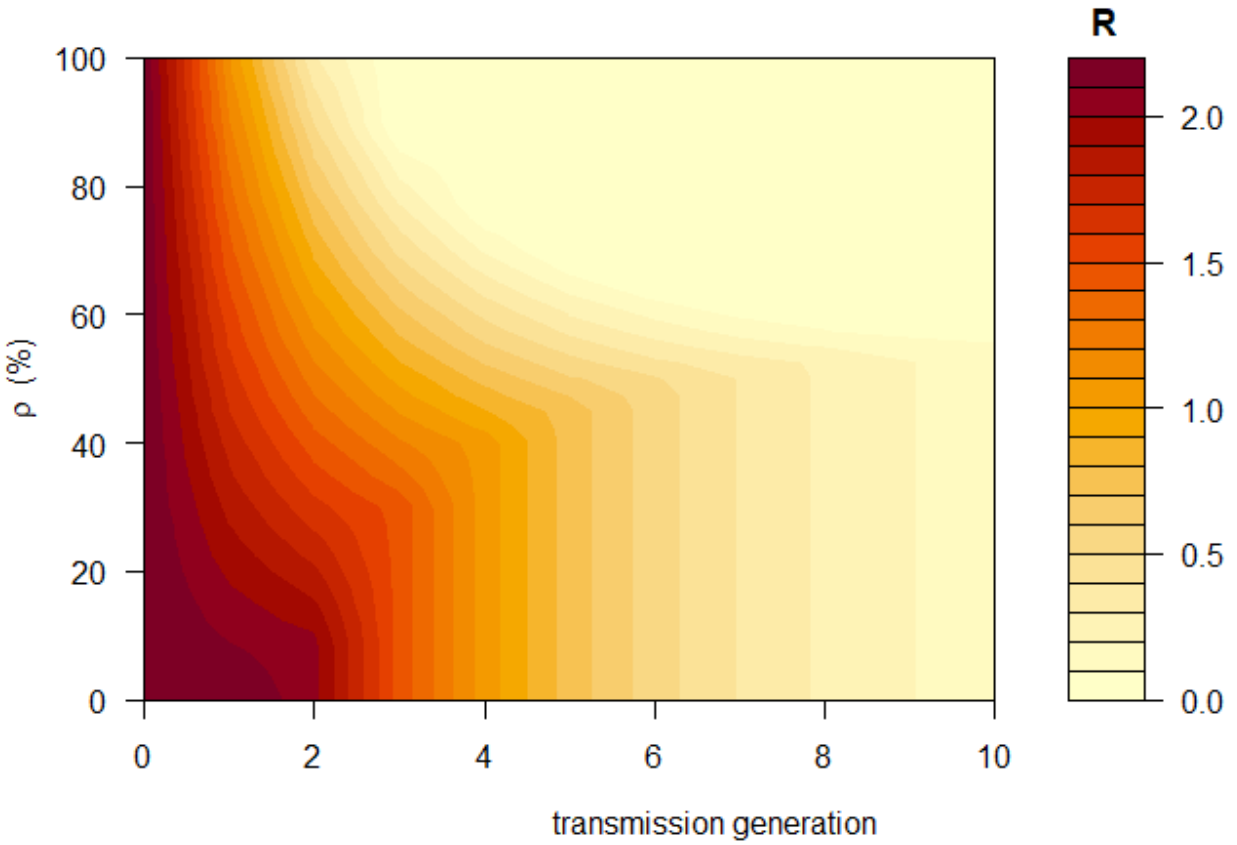
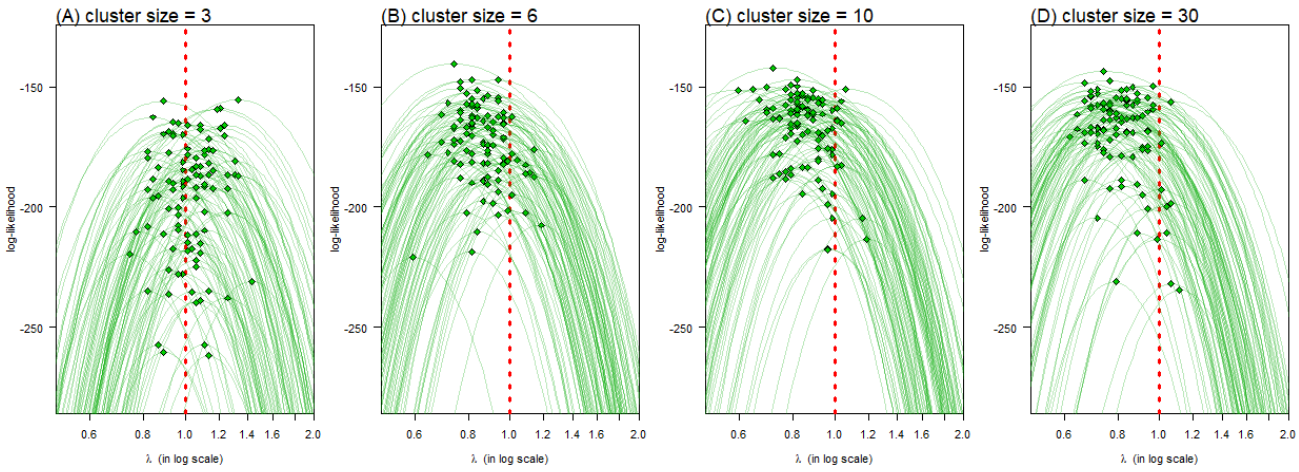


Figure 4.

The changing patterns of individual reproduction number (R) across the increasing transmission generations and the percentage of the reduction in SI due to the reduction in infectious period (ρ), see Section 2.5.1. For the initial (i.e., 0-th) transmission generation, the SI for the initial generation is fixed at 7.5 days, the latent period is fixed at 3.3 days, and the individual basic reproduction number (R_0) is fixed at 2.2.



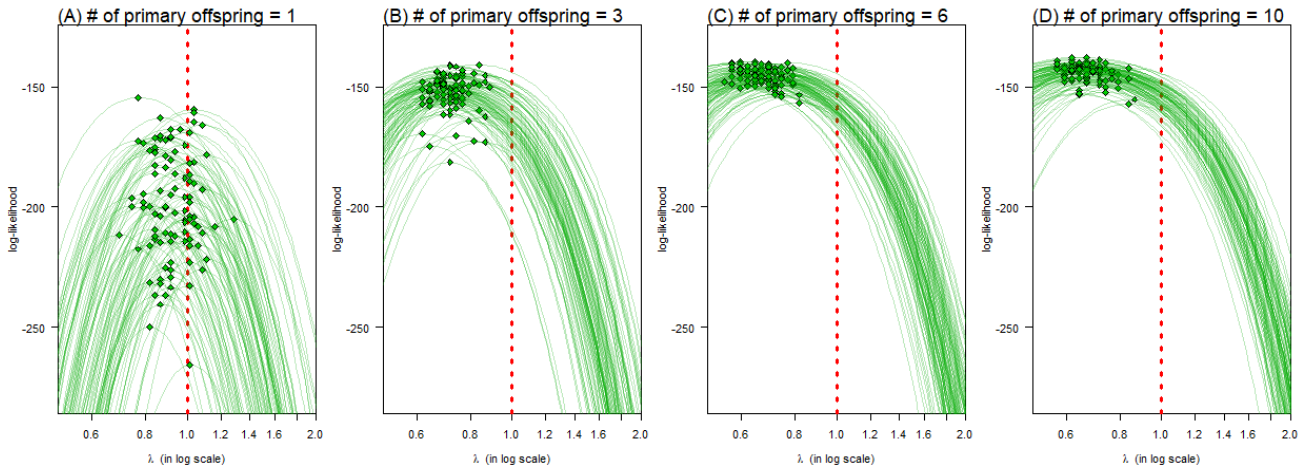
565

566 Figure 5.

567 The likelihood profiles of λ when the cluster size is 3 (A), 6 (B), 10 (C), or 30 (D). In each panel, the
568 green curves are the likelihood profiles of 100 set of samples (sample size of 30 for each set), and the
569 green dots are the maximum likelihood estimates of λ.

570

571



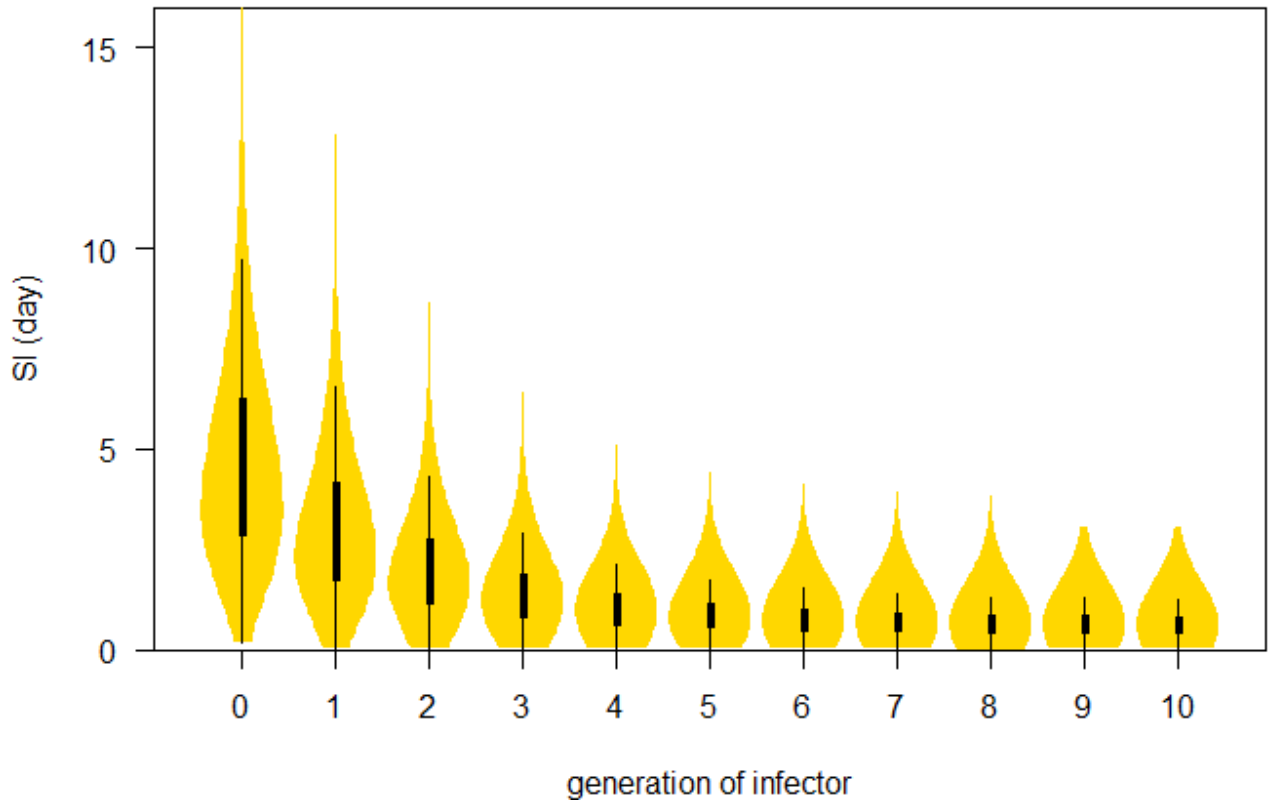
572

573 Figure 6.

574 The likelihood profiles of λ when the number of primary offsprings is 1 (A), 3 (B), 6 (C), or 10 (D).
575 In each panel, the green curves are the likelihood profiles of 100 set of samples (sample size of 30
576 for each set), and the green dots are the maximum likelihood estimates of λ.

577

578



579

580 Figure 7.

581 The distribution of SI of infector in each (cluster) generations. The gold area indicates the
582 distribution, the bold bars are the interquartile ranges (IQR), and the thin bars are the 95% centiles.

583

584

585 **Table**

586 Table 1.

587 Summary of the scale of change (λ) estimates (unit: per transmission generation). The shaded
588 estimates are considered as the main results.

SD of SI (σ)	Truncation	Distribution	scale of change (λ)	AICc
Large, i.e., SD = mean	No	Normal	0.66 (0.53, 0.82)	259.8
		Gumbel	0.78 (0.55, 1.11)	242.3
		Gamma	0.77 (0.51, 1.16)	237.7
	Yes	Normal	0.65 (0.53, 0.82)	224.7
		Gumbel	0.76 (0.52, 1.11)	212.5
		Gamma	0.72 (0.45, 1.16)	212.1
Moderate, i.e., SD^2 = mean	No	Normal	0.79 (0.66, 0.95)	275.2
		Gumbel	0.86 (0.74, 0.99)	252.2
		Gamma	0.87 (0.72, 1.05)	242.5
	Yes	Normal	0.69 (0.55, 0.87)	228.3
		Gumbel	0.74 (0.61, 0.91)	206.9
		Gamma	0.72 (0.54, 0.96)	200.6
Small, i.e., SD = 1	No	Normal	0.92 (0.86, 1.00)	552.6
		Gumbel	0.82 (0.81, 0.83)	6825.6
		Gamma	0.88 (0.83, 0.92)	634.0
	Yes	Normal	0.82 (0.73, 0.92)	452.2
		Gumbel	0.81 (0.80, 0.82)	5357.1
		Gamma	0.78 (0.73, 0.85)	494.8
crude estimate			1.00 (0.57, 1.43)	none

589

590

591