Title: The time-varying serial interval of the coronavirus disease (COVID-19) and its gender-specific difference: A data-driven analysis using public surveillance data in Hong Kong and Shenzhen, China from January 10 to February 15, 2020


Shi Zhao¹², Peihua Cao³, Marc KC Chong¹², Daozhou Gao⁴, Yijun Lou⁵, Jinjun Ran⁶, Kai Wang⁷, Weiming Wang⁸⁹, Lin Yang⁹, Daihai He⁵*, and Maggie H Wang¹²

¹ JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong, China
² Shenzhen Research Institute of Chinese University of Hong Kong, Shenzhen, China
³ Clinical Research Centre, Zhujiang Hospital, Southern Medical University, Guangzhou, China
⁴ Department of Mathematics, Shanghai Normal University, Shanghai, China
⁵ Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, China
⁶ School of Public Health, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong, China
⁷ Department of Medical Engineering and Technology, Xinjiang Medical University, Urumqi, 830011, China
⁸ School of Mathematics and Statistics, Huaiyin Normal University, Huaiian, China
⁹ School of Nursing, Hong Kong Polytechnic University, Hong Kong, China

* Correspondence to: zhaoshi.cmsa@gmail.com (SZ), weimingwang2003@163.com (WW), and daihai.he@polyu.edu.hk (DH)

Word count 564

Email address of all authors
SZ: zhaoshi.cmsa@gmail.com
PC: cphcc@smu.edu.cn
MKCC: marc@cuhk.edu.hk
DG: dzgao@shnu.edu.cn
YL: yijun.lou@polyu.edu.hk
JR: jimran@connect.hku.hk
KW: wangkaimath@sina.com
WW: weimingwang2003@163.com
LY: l.yang@polyu.edu.hk
DH: daihai.he@polyu.edu.hk
MHW: maggiew@cuhk.edu.hk

Keywords: coronavirus disease 2019; mitigation; serial interval; statistical analysis; gender.
Main text

To the Editor

An outbreak of coronavirus disease (COVID-19), started in Wuhan, China in the end of 2019 [1], has now reached 40 countries and poses huge threat to global public health and economic [2]. Given the risk of human-to-human transmission, the serial interval (SI), which refers to the time interval from symptom onset of a primary case (i.e., infector) to that of a secondary case (i.e., infectee) [3], is the other essential quantity besides the basic reproduction number to drive the spreading speed.

We examine the publicly available materials and collect the records of COVID-19 transmission events in two neighboring large cities, Hong Kong [4] and Shenzhen [5], in south China from January 10 to February 15, 2020 and extract the SI data. We identify 48 transmission events including 21 in Hong Kong and 27 in Shenzhen, among which 40 events contain the gender information of the primary cases. The last onset date of the primary cases among all collected transmission events is February 2, 2020.

To explore the temporal patterns and the gender-specific difference of SI, we adopt two regression models as follows.

- Model (1): log-linear form for the percentage change, \( E[\ln(\text{SI}_{i,t})] = \alpha_1 G_i + \alpha_2 t + \alpha_0; \)
  and
- Model (2): linear form for the unit change, \( E[\text{SI}_{i,t}] = \beta_1 G_i + \beta_2 t + \beta_0. \)

The \( E[\cdot] \) is the expectation, and the \( \alpha \) and \( \beta \) are the regression coefficients. The SI\(_{i,t}\) represents the SI of the \( i \)-th primary case whose onset date is the \( t \)-th day. The \( G_i \) denotes the gender of the \( i \)-th primary case. Hence, the \( \exp(\alpha_2) - 1 \) quantifies the percentage change, and \( \beta_2 \) quantifies the unit change (day) in the SI, namely change per day in the calendar date. The gender-specific difference can be interpreted similarly. We fit both models via the standard least square approach.

In Fig 1, we find that SI had been decreasing by 0.4 (95%CI: 0.1–0.7) per day, or 6.2% (95%CI: 0.4–11.6%) in percentage, from January 10 to February 2 in Hong Kong and Shenzhen. The Pearson correlation coefficient between the SI and calendar date is estimated at −0.37 with \( p \)-value < 0.01. The SI of male primary cases is 3.5 days (95%CI: 1.2–5.7) shorter than that of a female primary case, or 49.7% (95%CI: 15.3–70.1%) less in percentage.

To verify, we additionally conduct Cox proportional hazard modelling analysis using similar formula as in models (1) and (2), and calculate the hazard ratio estimates. We find the
association between SI and calendar date as well as gender-specific difference hold consistently and significantly.

The shortening in SI over time is likely due to the strengthening of the public health control measures. The contact tracing and timely isolation of confirmed COVID-19 infections could lead to shorter observed SI due to right censoring ‘bias’ [6, 7]. As such, we call the observed SI under the effects of control measures the effective SI, which has a mean 5.2 days from our dataset. This appears slightly but not significantly shorter than the previous estimated ‘intrinsic’ SI with a mean 7.5 days [1]. The mechanism behind the gender difference remains uncovered, but may be partly due to that male cases are more severe than female cases (“officials recorded a 2.8% fatality rate for male patients versus 1.7% for women” [8]). In this work, we report two findings in the SI of COVID-19 in Hong Kong and Shenzhen, and their implication warrants further investigation.

Declarations

Ethics approval and consent to participate
The data were collected via public domain, and thus neither ethical approval nor individual consent was not applicable.

Availability of materials
All data used in this work were publicly available via [4, 5], and the key R code was attached as a supplementary file.

Consent for publication
Not applicable.

Funding
DH was supported by General Research Fund (Grant Number 15205119) of the Research Grants Council (RGC) of Hong Kong, China. WW was supported by National Natural Science Foundation of China (Grant Number 61672013) and Huaian Key Laboratory for Infectious Diseases Control and Prevention (Grant Number HAP201704), Huaian, Jiangsu, China.

Acknowledgements
The author acknowledged the assists by Cindy Y Tian from the Chinese University of Hong Kong to fix the reference list.

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
The authors declared no competing interests.

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

References
4. The collection of Press Releases by the Centre for Health Protection (CHP) of Hong Kong. [https://www.chp.gov.hk/en/media/116/index.html]
Figure 1.

The observed (dots and bars) and fitted (curves) serial interval (SI) of COVID-19. The results of model (1) are shown in panel (a), and those of model (2) are shown in panel (b). In both panels, the red represents the female primary cases, and blue represents the male primary cases. The dots are the observed (or median) SI, and the bars are the ranges of SI for multiple primary cases. The bold curves are the fitting results, and the dashed curves are the 95% CIs.
Supplementary Material

S1 Key R code for analysis

```r
#
require(readxl)
library(survival)
#
SI.data = read_excel(path = 'comb_data.xlsx', sheet = 'comb_data', na = 'NA')
SI.data = as.data.frame(SI.data)
SI.data = SI.data[,1:9]
SI.data$Infector.date.lwr <- as.Date(SI.data$Infector.date.lwr,format ='%m/%d/%Y')
SI.data$Infector.date.upr <- as.Date(SI.data$Infector.date.upr,format ='%m/%d/%Y')
SI.data$Infectee.date <- as.Date(SI.data$Infectee.date,format ='%m/%d/%Y')
SI.data$mid.Infectee.date.index = (as.numeric(SI.data$Infector.date.lwr
- as.Date('2019-12-31')) + as.numeric(SI.data$Infector.date.upr -as.Date('2019-12-31'))) /2

lwr.time.array = NULL
upr.time.array = NULL
surv.code.array = NULL
for (i in 1:nrow(SI.data)) {#       i = 1
  temp.data = SI.data[i,]
  temp.lwr = as.numeric(temp.data$Infectee.date - temp.data$Infector.date.upr)
  temp.upr = as.numeric(temp.data$Infectee.date - temp.data$Infector.date.lwr)
  temp.lwr = ifelse(temp.lwr <=0, 0.5, temp.lwr)
  temp.upr = ifelse(temp.upr <temp.lwr, temp.lwr, temp.upr)
  temp.code = ifelse(temp.upr > temp.lwr, 3, 1)
  lwr.time.array = c(lwr.time.array, temp.lwr)
}
```
upr.time.array = c(upr.time.array, temp.upr)
surv.code.array = c(surv.code.array, temp.code)
}
SI.data$lwr.SI = lwr.time.array
SI.data$upr.SI = upr.time.array
SI.data$mid.SI = (SI.data$lwr.SI + SI.data$upr.SI) / 2
SI.data$surv.code = surv.code.array

#

sel.data = SI.data
#
cor.test(sel.data$mid.Infectee.date.index, sel.data$mid.SI, method = 'p')
#
simple.lm = lm(log(mid.SI) ~ mid.Infectee.date.index +Infector.gender, data = sel.data)
simple.lm = lm(c(mid.SI) ~ mid.Infectee.date.index +Infector.gender, data = sel.data)
#
surv.obj = Surv(time = sel.data$mid.SI, event = rep(1, nrow(sel.data)))

cx.mod = coxph(surv.obj ~ mid.Infectee.date.index +Infector.gender, data = sel.data)
<table>
<thead>
<tr>
<th>Infector.date.lwr</th>
<th>Infector.date.upr</th>
<th>Infector.gender</th>
<th>Infector.age</th>
<th>Infectee.date</th>
<th>Infectee.gender</th>
<th>Infectee.age</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/19/2020</td>
<td>1/20/2020</td>
<td>M</td>
<td>73</td>
<td>1/22/2020</td>
<td>F</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>1/23/2020</td>
<td>1/24/2020</td>
<td>M</td>
<td>75</td>
<td>1/26/2020</td>
<td>F</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>1/24/2020</td>
<td>1/25/2020</td>
<td>M</td>
<td>68</td>
<td>1/26/2020</td>
<td>F</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>1/26/2020</td>
<td>1/27/2020</td>
<td>M</td>
<td>68</td>
<td>1/28/2020</td>
<td>F</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>1/29/2020</td>
<td>1/30/2020</td>
<td>M</td>
<td>68</td>
<td>1/31/2020</td>
<td>F</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>1/30/2020</td>
<td>1/31/2020</td>
<td>M</td>
<td>68</td>
<td>2/1/2020</td>
<td>F</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>1/31/2020</td>
<td>2/1/2020</td>
<td>M</td>
<td>68</td>
<td>2/2/2020</td>
<td>F</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>2/5/2020</td>
<td>2/6/2020</td>
<td>M</td>
<td>68</td>
<td>2/7/2020</td>
<td>F</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>2/6/2020</td>
<td>2/7/2020</td>
<td>M</td>
<td>68</td>
<td>2/8/2020</td>
<td>F</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2/7/2020</td>
<td>2/8/2020</td>
<td>M</td>
<td>68</td>
<td>2/9/2020</td>
<td>F</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2/8/2020</td>
<td>2/9/2020</td>
<td>M</td>
<td>68</td>
<td>2/10/2020</td>
<td>F</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>