Serial Interval Distribution of SARS-CoV-2 Infection in Brazil

Supplemental Material

Carlos A. Prete Jr[1], Lewis Buss[2], Amy Dighe[3], Victor Bertollo Porto[4], Darlan da Silva Candido[5], Fábio Ghilardi[2], Oliver G. Pybus[5], Wanderson K. de Oliveira[4], Júlio H. R. Croda[4], Ester C. Sabino[2], Nuno Rodrigues Faria[2,3,5], Christl A. Donnelly[2,6], Vítor H. Nascimento[1].

1. Department of Electronic Systems Engineering, University of São Paulo, Brazil
2. Instituto de Medicina Tropical, University of São Paulo, Brazil
3. MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, United Kingdom
4. Secretaria de Vigilância em Saúde, Ministério da Saúde, Brazil
5. Department of Zoology, University of Oxford, United Kingdom
6. Department of Statistics, University of Oxford, United Kingdom

Derivation of the serial interval distribution

In order to derive the serial interval probability distribution, we decompose the instant of first symptoms in terms of simpler random variables and use this expression to derive the serial interval in terms of these random variables. Then, we use the serial intervals measured from our data to choose a distribution for these random variables and consequently obtain the serial interval distribution.

The instant of first symptom onset $t_{\mathrm{sym},1}$ and $t_{\mathrm{sym},2}$ for the primary and secondary case respectively is given by the sum of the instant of infection t_{inf}, the latent period Δt_{lat} (that is, the time the infection takes to become transmissible) and the interval Δt_{int} between the end of latent period and the instant of first symptoms:

$$\begin{cases} t_{\mathrm{sym},1} = t_{\mathrm{inf},1} + \Delta t_{\mathrm{lat},1} + \Delta t_{\mathrm{int},1} \\ t_{\mathrm{sym},2} = t_{\mathrm{inf},2} + \Delta t_{\mathrm{lat},2} + \Delta t_{\mathrm{int},2} \end{cases}.$$

Because only the day of symptom onset was available, the exact time of first symptoms is unknown. For this reason, denoting as $\lfloor x \rfloor$ the largest integer less than or equal to x (the floor operator), the measured date of first symptoms for the primary and secondary patients is respectively $\lfloor t_{\mathrm{sym},1} \rfloor$ and $\lfloor t_{\mathrm{sym},2} \rfloor$. Hence, the measured serial interval ΔD_{sym}, defined as the difference of dates of symptoms, is given by

$$\Delta D_{\mathrm{sym}} = \lfloor t_{\mathrm{sym},2} \rfloor - \lfloor t_{\mathrm{sym},1} \rfloor = \lfloor t_{\mathrm{inf},2} + \Delta t_{\mathrm{lat},2} + \Delta t_{\mathrm{int},2} \rfloor - \lfloor t_{\mathrm{inf},1} + \Delta t_{\mathrm{lat},1} + \Delta t_{\mathrm{int},1} \rfloor.$$

Defining Δt_{tra} as the time the primary patient takes to infect the secondary patient counting from the instant where the disease can be transmitted, the instant of infection of the secondary patient can be written as $t_{\mathrm{inf},2} = t_{\mathrm{inf},1} + \Delta t_{\mathrm{inc},1} + \Delta t_{\mathrm{tra}}$, which can be substituted in the expression of ΔD_{sym}, yielding

$$\Delta D_{\mathrm{sym}} = \lfloor t_{\mathrm{inf},1} + \Delta t_{\mathrm{lat},1} + \Delta t_{\mathrm{tra}} + \Delta t_{\mathrm{lat},2} + \Delta t_{\mathrm{int},2} \rfloor - \lfloor t_{\mathrm{inf},1} + \Delta t_{\mathrm{lat},1} + \Delta t_{\mathrm{int},1} \rfloor.$$

We modelled $t_{\mathrm{inf},1}$ as a uniform random variable in the interval $[0, 1]$ because ΔD_{sym} is only affected by the non-integer part of $t_{\mathrm{inf},1}$. In order to assign distributions for Δt_{tra}, $\Delta t_{\mathrm{lat},k}$ and $\Delta t_{\mathrm{int},k}$ $(k = 1,2)$, we assumed that each of these variables may follow one of the following distributions: Log-normal, Gamma, Poisson, Exponential, Weibull, Chi-Squared and zero (that is, a deterministic variable that is equal to zero). Then, considering that $\Delta t_{\mathrm{lat},1}$ and $\Delta t_{\mathrm{int},1}$ have the same distribution as $\Delta t_{\mathrm{lat},2}$ and $\Delta t_{\mathrm{int},2}$ respectively, we estimated the distribution of ΔD_{sym} for all $7^3 = 343$ possible distributions of Δt_{tra}, $\Delta t_{\mathrm{lat},k}$ and $\Delta t_{\mathrm{int},k}$ $(k = 1,2)$ through Monte Carlo simulation using 10^8 realizations and choosing the parameters of the distributions using the maximum likelihood estimator. Hence, we generated 343 different models for ΔD_{sym} and picked the model that yields the best Akaike Information Criterion (AIC) when fitted to our data. Note that even though our method returns a distribution for ΔD_{sym}, the models for Δt_{tra}, $\Delta t_{\mathrm{lat},k}$ and $\Delta t_{\mathrm{int},k}$ obtained by our method cannot be interpreted as estimates for the distributions of these variables, since different models for Δt_{tra}, $\Delta t_{\mathrm{lat},k}$ and $\Delta t_{\mathrm{int},k}$ may yield the same distribution for ΔD_{sym}.

The best model we obtained uses $\Delta t_{\mathrm{tra}} = 0$ and assigns a chi-squared distribution to $\Delta t_{\mathrm{lat},k}$ and $\Delta t_{\mathrm{int},k}$ $(k = 1, 2)$ with means $a = 3.03$ and $b = 0.95$ days respectively. Hence, our model for ΔD_{sym} is

$$\Delta D_{\mathrm{sym}} = \left\lfloor t_{\mathrm{inf},1} + \Delta t_{\mathrm{lat},1} + \Delta t_{\mathrm{lat},2} + \Delta t_{\mathrm{int},2} \right\rfloor - \left\lfloor t_{\mathrm{inf},1} + \Delta t_{\mathrm{lat},1} + \Delta t_{\mathrm{int},1} \right\rfloor,$$

where $t_{\mathrm{inf},1}$ is a uniform random variable in the interval [0,1], $\Delta t_{\mathrm{lat},1}$ and $\Delta t_{\mathrm{lat},2}$ are independent chi-squared random variables of mean $a = 3.03$ and $\Delta t_{\mathrm{int},1}$ and $\Delta t_{\mathrm{int},2}$ are independent chi-squared random variables of mean $b = 0.95$. This serial interval model has only two parameters (a and b), which is the same number of parameters used in many other

serial interval models, as Gaussian distribution, log-normal distribution and Weibull distribution [1].

It is worth noting that if the uncertainty on ΔD_{sym} caused by the specification of the date of first symptoms instead of the exact instant of first symptoms is ignored, the floor operator is not needed, thus ΔD_{sym} is simply modelled as the difference of two chi-squared random variables X and Y as $\Delta D_{\mathrm{sym}} = X - Y$, where $X = \Delta t_{\mathrm{lat},2} + \Delta t_{\mathrm{int},2}$ and $Y = \Delta t_{\mathrm{int},1}$. In this case, the mean measured serial interval is simply given by a, the expected value of $\Delta t_{\mathrm{lat},k}$.

In order to assess the reliability of the estimated parameters, we resampled the measured serial intervals using bootstrapping with 1,000 bootstrap samples, and we fitted our model to each resampled dataset, obtaining 1,000 values for the parameters a and b. The estimated mean (standard deviation) of a and b are respectively 2.99 (0.37) and 0.94 (0.47), and their 95% confidence intervals are respectively $[2.26, 3.73]$ and $[0.19, 2.03]$.

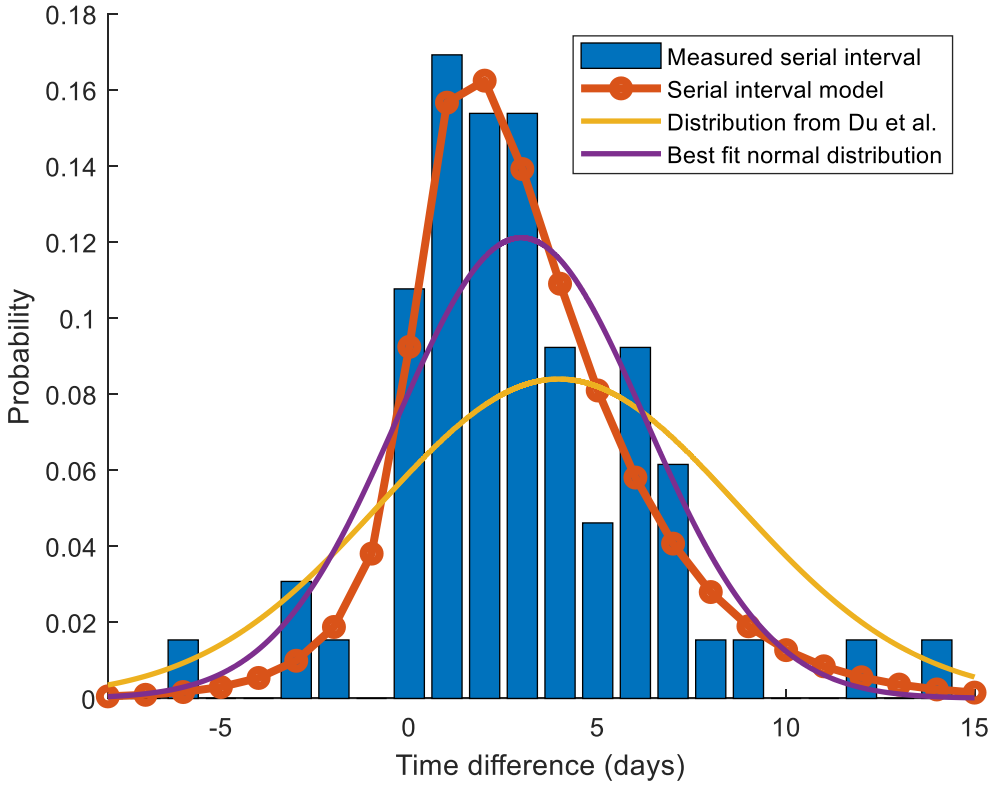
Supplemental Table 1 compares the mean, median, standard deviation probability of negative serial intervals between the observed data and the model. It also contains the mean, variance and 95% confidence interval of these statistics for our model, which were estimated using the data resampled with bootstrapping. A goodness-of-fit test using 1,000 bootstrap samples was also performed, following Szucs [2] and Stute et al. [3]. The Kolmogorov-Smirnov distance between the empirical distribution and the distribution obtained with the estimated parameters was 0.267, while the threshold for a significance level of 0.05 was 0.685.

	Mean	Median	Standard deviation	Probability of negative serial intervals
Measured serial intervals	2.97	3	3.29	6.15%
Model for serial intervals	3.03	3	3.16	7.89%
Mean (bootstrapping)	2.99	2.53	3.13	7.78%
Standard deviation (bootstrapping)	0.37	0.51	0.31	3.47%
95% confidence interval (bootstrapping)	[2.26, 3.73]	[2, 3]	[2.55, 3.80]	[1.72%, 15.29%]

Supplemental Table 1. Comparison of mean, median, standard deviation and probability of negative serial intervals computed directly from the serial intervals measured from data and from our model for the serial intervals. This table also contains the mean, standard deviation and 95% confidence interval for each statistic, which were obtained from our serial interval model using bootstrapping.

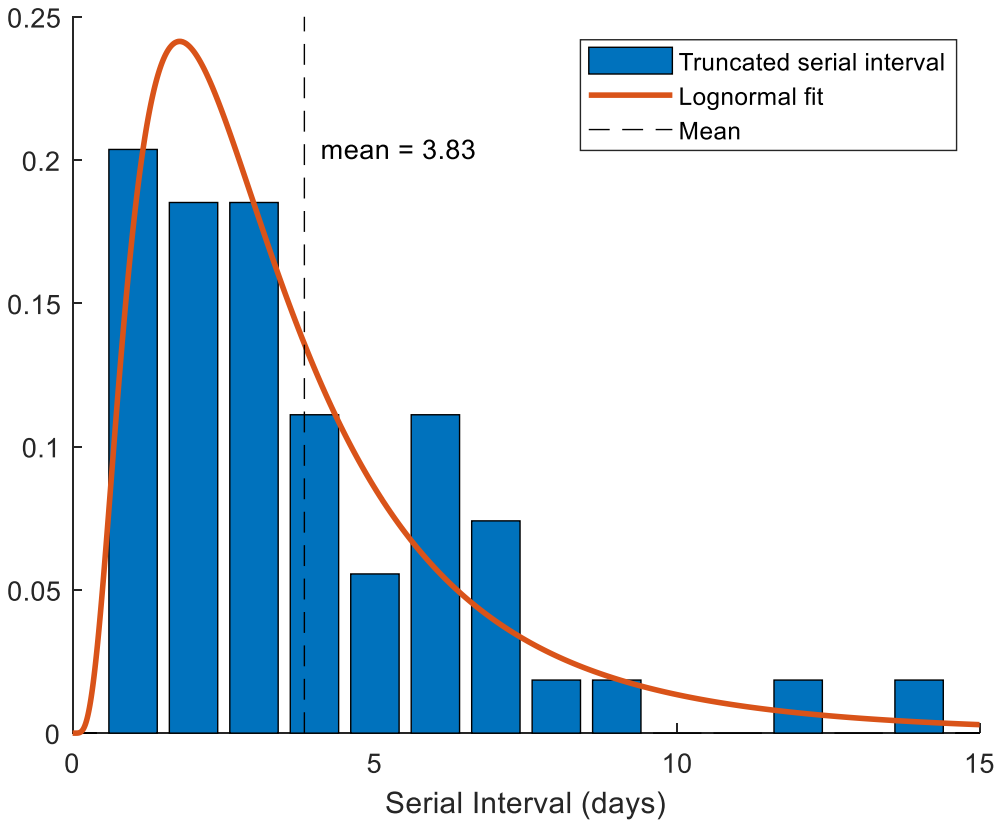
Alternative distributional approaches

Du et al. [4] present a large data set of transmission pairs in China extracted from media and government reports. 12% of their infection pairs reported negative serial intervals. They find that their data are best fit by a normal distribution. For comparison we have fit the observed Brazilian serial intervals with a normal distribution using maximum likelihood estimation, as shown in **Supplemental Figure 1**, with the distribution from Du et al. [4] superimposed. The Gaussian distribution fitted to our data has mean 2.97 days and standard deviation 3.29 days.



Supplemental Figure 1. Brazilian serial interval data with the best fit normal distribution, our proposed serial interval distribution and the distribution taken from Du et al. [1] superimposed for comparison.

It is conceivable that very short, or negative serial intervals represent clusters in which a third (unreported) individual was the true source, making these values less reliable. Other authors [5] have dealt with this issue by censoring the observed observations to only positive values. Taking this approach, we removed four negative valued and seven zero-valued serial intervals leaving 54 observations. The mean of these truncated values was 3.83 with a standard deviation of 2.8 and median of 3 days. We tested normal, lognormal, Gamma and Weibull distributions. The best fitting distribution by the Akaike information criterion was lognormal with meanlog of 1.09 and sdlog 0.72. See **Supplemental Figure 2**.



Supplemental Figure 2. Serial interval data reported to the Brazilian Ministry of Health truncated to include only positive values. This is overlaid with the best fit lognormal distribution.

References

1. Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. International Journal of Infectious Diseases, Volume 93, 2020, Available at: http://medrxiv.org/lookup/doi/10.1101/2020.02.03.20019497. Accessed 24 March 2020

2. Szűcs, G. Parametric bootstrap tests for continuous and discrete distributions. Metrika 67, 63–81 (2007).

3. Stute, W., Manteiga, W. G. & Quindimil, M. P. Bootstrap based goodness-of-fit-tests. Metrika 40, 243–256 (1993).

4. Du Z, Xu X, Wu Y, Wang L, Cowling BJ, Meyers L. Serial Interval of COVID-19 among Publicly Reported Confirmed Cases. Emerg Infect Dis. 2020;26(6):1341-1343. https://dx.doi.org/10.3201/eid2606.200357

5. Zhang J, Litvinova M, Wang W, et al. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. The Lancet Infectious Diseases 2020; :S1473309920302309.