Biases in early-outbreak estimates of epidemiological delay distributions: applications to COVID-19 outbreak

Abstract

1 Introduction

Since the emergence of the novel coronavirus disease (COVID-19), a significant amount of research has been put into characterizing its course of infection. This is done by measuring individual-level delays between key epidemiological events — either within an infected individual or between transmission pairs — and estimating their population-level probability distributions. Key distributions for understanding the spread of COVID-19 include:

1. Incubation period distribution: time between infection and symptom onset (Backer et al., 2020; Li et al., 2020; Linton et al., 2020; Tian et al., 2020)

2. Serial interval distribution: time between symptom onset of an infector and an infectee (Du et al., 2020; Nishiura et al., 2020; Zhao et al., 2020,?)

3. Generation interval distribution: time between infection of an infector and an infectee (Ganyani et al., 2020)

The inferred delay distributions then allow us to estimate the epidemic potential of COVID-19 and assess the effectiveness of intervention strategies.

Measuring a delay between two epidemiological events depends on having observed both events. A delay between two events cannot be measured if the second event has not occurred or has not been observed yet. Here, we show that this dependency can systematically bias the estimate of a delay distribution if it is not explicitly taken into account; this bias applies to *all* epidemiological delay distributions. We compare two approaches for correcting the bias and apply them to evaluate the amount of bias present in the early-outbreak estimate of the mean incubation period.

2 Theoretical framework

In order to characterize how the observed delay distributions between two epidemiological events vary over time, we begin by defining the epidemiological delays from a cohort-based perspective. A cohort at time s consists of all infected individuals whose first epidemiological event of interest occurred at time s. For example, if we are interested in measuring the duration of symptoms, a cohort at time s consists of all individuals who became symptomatic at time s, regardless of when they were infected.

Cohort-based delay distributions are always subject to right-censoring. In order to measure a delay between two epidemiological events, both events must occur before the time of measurement t, and therefore, delays that are longer than $t - s$ cannot be observed for a cohort at time $s < t$. Then, the cohort delay distribution $c_s(\tau|t)$ can be expressed as a truncated distribution:

$$c_s(\tau|t) = \frac{f_s(\tau)}{F_s(t-s)}, \tag{1}$$

where $f_s(\tau)$ is the true delay distribution for cohort s, $F_s(\tau)$ is the corresponding cumulative distribution function, and $0 \leq \tau \leq t - s$. While delay distributions measured within an infected individual (e.g., incubation period) may not vary much over the course of an epidemic,

those measured between infected individuals (e.g., generation interval) vary due to changes in the susceptible population. We note that within-individual delay distributions that depend on intervention strategies (e.g., time between symptom onset and hospitalization) can also change over time.

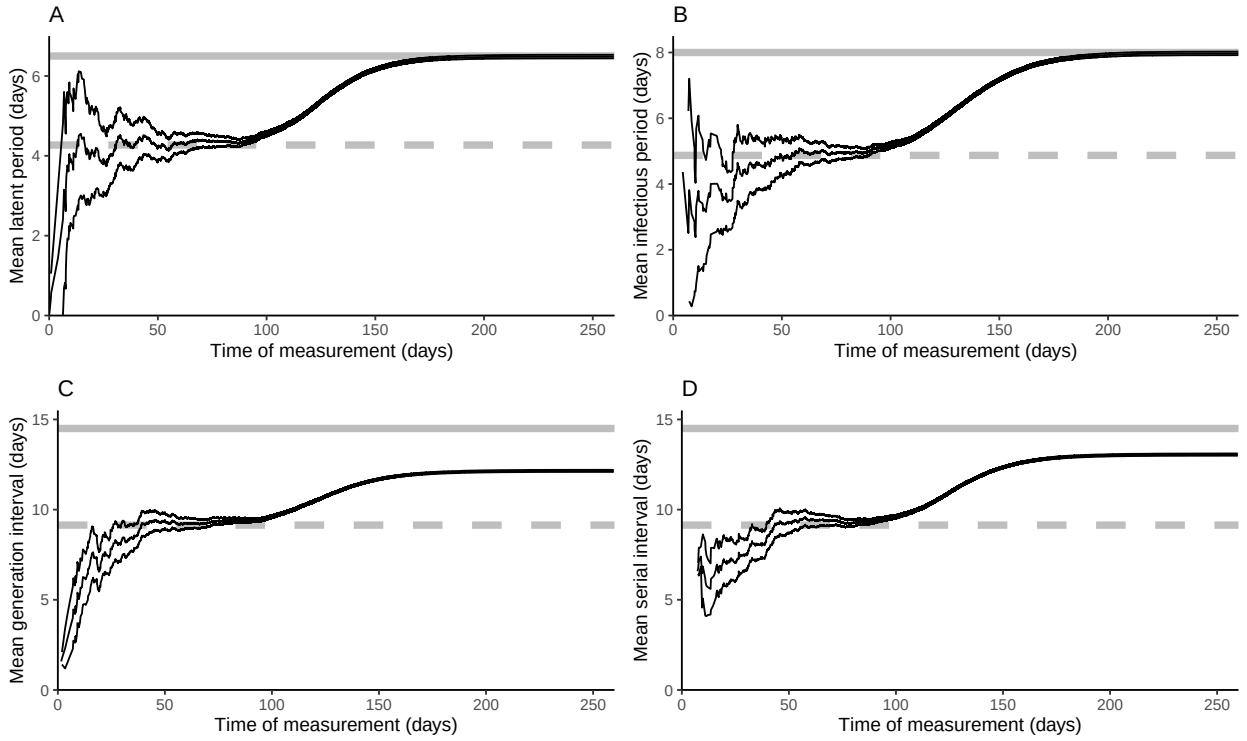

Figure 1: **Observed means of epidemiological delay distributions over time.** Changes in the observed mean latent period (A), infectious period (B), generation interval (C), and serial interval (D) over the course of an epidemic. Black solid lines represent the observed means and associated 95% confidence intervals, which are calculated by taking all samples until the time of measurement t. Gray solid lines represent the true mean. Gray dashed lines represent the expected mean during the exponential growth phase that we calculate using Equation (4). A stochastic SEIR model was simulated using a COVID-like parameters: $\mathcal{R}_0 = 2.5$, $1/\sigma = 6.5$ days, $1/\gamma = 8$ days, $N = 100000$, and $I(0) = 10$.

Typically, epidemiological delay distributions are estimated by using *all* available measured samples. Then, the observed delay distribution $f_{\text{obs}}(\tau|t)$, which takes into account all measured delays until time t, can be expressed as an average of the cohort delay distributions $c_s(\tau|t)$, weighted by the incidence of cohorts $i(s)$ and the probability that both epidemiological events of interest will occur between time s and t:

$$\begin{aligned} f_{\text{obs}}(\tau|t) &\propto \int_{-\infty}^{t-\tau} c_s(\tau|t) i(s) F_s(t-s) ds \\ &= \int_{-\infty}^{t-\tau} i(s) f_s(\tau) ds \end{aligned} \tag{2}$$

Early in an epidemic, the incidence of infection, and therefore the incidence of cohorts, is expected to grow exponentially at rate $r > 0$: $i(s) \propto \exp(rs)$. Assuming that the true delay distribution stays constant during this period ($f_s(\tau) = f(\tau)$), the observed delay distribution during the exponential growth phase $f_{\exp}(\tau|t)$ is equivalent to the true distribution weighted by the inverse of the exponential growth rate (Britton and Scalia Tomba, 2019):

$$\begin{aligned} f_{\exp}(\tau|t) &\propto f(\tau) \int_{-\infty}^{t-\tau} \exp(rs) ds \\ &\propto f(\tau) \exp(-r\tau) \end{aligned} \tag{3}$$

In other words, for fast-growing epidemics (high r), there will be a stronger bias to observe shorter intervals. This bias applies to all epidemiological delay distributions.

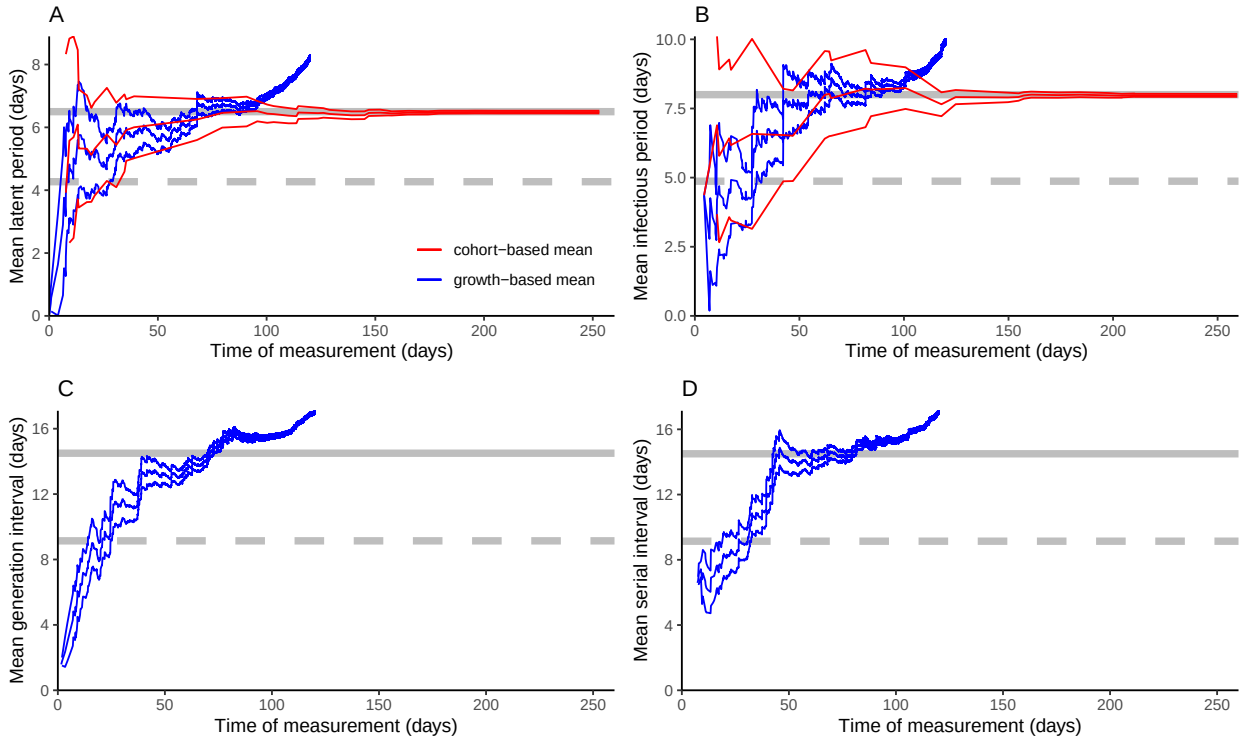

Figure 2: **Estimated means of epidemiological delay distributions over time.** Changes in the estimated mean latent period (A), infectious period (B), generation interval (C), and serial interval (D) over the course of an epidemic. Black solid lines represent the estimated mean and associated 95% confidence intervals at each time of measurement using the growth-based approach. Red solid lines represent the estimated mean and associated 95% confidence intervals at each time of measurement using the cohort-based approach. Gray solid lines represent the true mean. Gray dashed lines represent the expected observed mean during the exponential growth phase calculated using Equation (4). A stochastic SEIR model was simulated using the following parameters: $\mathcal{R}_0 = 2.5$, $1/\sigma = 6.5$ days, $1/\gamma = 8$ days, $N = 50000$, and $I(0) = 10$.

We compare how the observed mean latent period, infectious period, generation interval, and serial interval (assuming that the latent period is equivalent to the incubation period) change over time using an individual-based stochastic simulation of the SEIR model (Fig. 1). The stochastic simulation confirms the bias: the observed mean delay during the exponential growth phase matches the expected mean that we calculate using Equation (4). The amount of bias decreases as the epidemic progresses. The observed mean latent and infectious periods become unbiased eventually but not until much later in the epidemic.

The observed mean generation and serial intervals remain biased even at the end of the epidemic because the realized generation intervals become shorter due to susceptible depletion (Champredon and Dushoff, 2015). The susceptible depletion effect can even occur at a finer, local scale (cf. Park et al. (2019)). The observed mean serial interval is slightly higher than the observed mean generation interval near the end of an epidemic because an infected individual with a short latent period (and shorter incubation period) is more likely infect others by transmitting faster (therefore, shorter generation interval) during the susceptible depletion phase; therefore, infectors are more likely to have shorter latent/incubation periods than their infectees. We expect intervention strategies to have similar effects on generation intervals because faster transmission events (i.e., shorter generation intervals) will be more likely to evade intervention.

Equation (4) suggests a seemingly straightforward way of correcting the bias – by weighting the observed distribution by the exponential growth rate:

$$f(\tau) = f_{\text{exp}}(\tau|t)\exp(r\tau). \tag{4}$$

Similar forms have been suggested by other studies (Britton and Scalia Tomba, 2019; Park et al., 2019) and have been applied in estimating epidemiological delay distributions during the COVID-19 outbreak (Nishiura et al., 2020; Linton et al., 2020); however, our simulations show that the growth-based approach does not work very well for estimating the mean delay (Fig. 2). Although the estimated mean delay is consistent with true mean, the estimates are unstable and the associated confidence intervals are narrow. Once the exponential growth phase is over, the growth-based approach is no longer viable as it will overestimate the mean.

Alternatively, we can account for the bias by ensuring that the right-censoring does not exist in the sample: instead of using all samples that have been collected until the time of measurement, we can limit our samples to cohort $u < t$ such that both epidemiological events of interest have been observed for all individuals within cohorts $s < u$. This approach provides an unbiased estimate of the mean delay throughout the epidemic with appropriately wide confidence intervals that contain the true value (Fig. 2); conversely, the proportion of individuals that have not completed the second event within each cohort will be indicative of the amount of bias present in the estimate. This approach cannot be applied to generation or serial intervals because we don't know how many individuals each person infected exactly (and therefore we cannot evaluate whether the right-censoring exists within a cohort). Nonetheless, likelihood-based methods that explicitly account for the right-censoring can be still applied to estimate the generation interval distribution (Park et al., 2019).

3 Applications: incubation period distribution of COVID-19

Here, we revisit the mean incubation period estimated by Backer et al. (2020) and assess the degree of potential bias in their estimate. Since they relied on early traveler data who traveled from Wuhan between January 2–23, during which the epidemic was likely to have been expanding exponentially, the number of infected travelers from Wuhan was likely to have been increasing exponentially as well. Therefore, their estimate of the mean *observed* incubation periods of 6.4 days (95% CI: 5.6–7.7 days) is likely to be biased due to right-censoring.

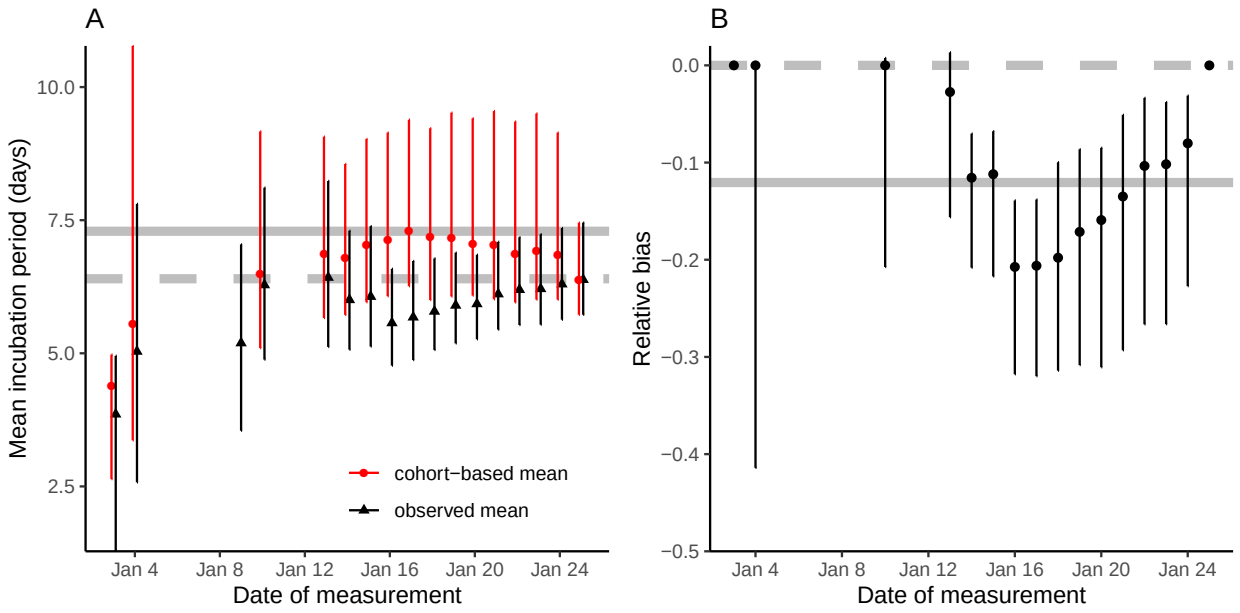

Figure 3: **Potential bias in the early-outbreak estimate of the mean incubation period.** (A) Comparisons of the observed mean incubation period using all available samples and the cohort-based mean incubation period. Gray solid line represents the growth-based mean incubation period. Gray dashed line represents the mean incubation period estimated by Backer et al. (2020). (B) Relative bias of the mean observed incubation period with respect to the cohort-based mean incubation period. Gray solid line represents relative bias with respect to the growth-based mean incubation period. Gray dashed line represents the $y = 0$ line. Relative bias is calculated as (observed mean incubation period)/(bias-corrected mean incubation period) − 1.

Fig. **??**A compares the observed mean incubation period, which uses all available measurements, with the cohort-based mean incubation period (see Methods for details of this section). Consistent with our simulations, the cohort-based means are generally higher and have wider confidence intervals (see January 13–24 in Fig. **??**A). For example, on January 24, the cohort-based mean incubation period is 6.9 days (95% CI: 6.0 – 9.1 days), whereas

the observed mean incubation period is 6.3 days (95% CI: 5.6 days – 7.4 days). The cohort-based mean also matches the growth-based mean: 7.3 days (95% CI: 6.2 – 9.9 days; see solid gray line in Fig. **??**A). On January 25, two estimates completely overlap because they both use all available samples; a sudden decrease in the cohort-based mean indicates the presence of right-censoring. Since Fig. **??**A compares the marginal posterior distributions of the means, it does not allow us to assess whether the naive mean is lower than the cohort-based adjusted mean – the overlapping confidence intervals do not imply that the differences are not statistically clear (Dushoff et al., 2019).

Fig. **??**B compares the relative bias of the observed means with respect to the cohort-based adjusted means. We find clear and consistent bias between January 14–24; the median estimates of the bias range from 8% to 20%. These estimates are also consistent with the amount of bias that we calculate using the growth-based means: 12% (95% CI: 6%–26%). Overall, our results indicate that the early-outbreak estimate of the mean incubation period of COVID-19 by Backer et al. (2020) was likely to have been biased.

4 Discussion

Estimating the time distributions underlying the spread of disease is crucial for predicting the course of an outbreak and controlling it. However, estimates of epidemiological delay distributions can be systematically biased during an ongoing epidemic due to right-censoring. Generation and serial intervals can be subject to additional bias due to susceptible depletion and are more difficult to estimate.

Knowing the incubation period of novel pathogens, such as COVID-19, is important for designing initial intervention strategies, including quarantine measures. For example, 14-day isolation is currently recommended for anyone who comes in close contacts with confirmed COVID-19 cases in many countries to reflect its estimated incubation period; nonetheless, a few studies have documented much longer incubation periods for COVID-19, ranging from 19 to 24 days (Bai et al., 2020; Guan et al., 2020). Our analysis suggests that early estimates of the mean incubation period may be biased; we recommend reassessing the effectiveness of measures that rely on this estimate.

We compared two approaches for correcting the right-censoring bias: growth-based approach and cohort-based approach. While the growth-based approach provides a simple, intuitive way of assessing the bias present in the estimate, it is unstable as it is overly sensitive to long intervals and assumes that the exponential growth rate is exactly known. Given that the exact period of exponential growth is difficult to determine (Ma et al., 2014), we recommend against using growth-based approaches. The cohort-based approach provides unbiased estimates throughout the course of an epidemic. In practice, using only a subset of available samples may not be ideal as it leads to less precise inference; we recommend using likelihood-based methods that explicitly account for right-censoring analogous to Equation (1). Nonetheless, comparing the cohort-based means and the observed means still provides a viable way of assessing whether estimates are biased.

As COVID-19 continues to appear in new countries, researchers will prioritize charac-

terizing differences in local patterns of spread. However, the observed epidemic patterns in these locations will be subject to stronger biases than those in previously established regions, such as China or South Korea. We strongly suggest future research to consider *all* sources of potential biases in the early-outbreak patterns of spread.

References

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

Bai, Y., L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, and M. Wang (2020). Presumed asymptomatic carrier transmission of COVID-19. *Jama*.

Britton, T. and G. Scalia Tomba (2019). Estimation in emerging epidemics: Biases and remedies. *Journal of the Royal Society Interface 16*(150), 20180670.

Champredon, D. and J. Dushoff (2015). Intrinsic and realized generation intervals in infectious-disease transmission. *Proceedings of the Royal Society B: Biological Sciences 282*(1821), 20152026.

Du, Z., X. Xu, Y. Wu, L. Wang, B. J. Cowling, and L. A. Meyers (2020). The serial interval of COVID-19 from publicly reported confirmed cases. *medRxiv*.

Dushoff, J., M. P. Kain, and B. M. Bolker (2019). I can see clearly now: Reinterpreting statistical significance. *Methods in Ecology and Evolution 10*(6), 756–759.

Ganyani, T., C. Kremer, D. Chen, A. Torneri, C. Faes, J. Wallinga, and N. Hens (2020). Estimating the generation interval for COVID-19 based on symptom onset data. *medRxiv*.

Guan, W.-j., Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. Hui, et al. (2020). Clinical characteristics of 2019 novel coronavirus infection in China. *MedRxiv*.

Li, Q., X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong, et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.

Linton, N. M., T. Kobayashi, Y. Yang, K. Hayashi, A. R. Akhmetzhanov, S.-m. Jung, B. Yuan, R. Kinoshita, and H. Nishiura (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine 9*(2), 538.

Ma, J., J. Dushoff, B. M. Bolker, and D. J. Earn (2014). Estimating initial epidemic growth rates. *Bulletin of mathematical biology 76*(1), 245–260.

Nishiura, H., N. M. Linton, and A. R. Akhmetzhanov (2020). Serial interval of novel coronavirus (COVID-19) infections. *International Journal of Infectious Diseases*.

Park, S. W., D. Champredon, and J. Dushoff (2019). Inferring generation-interval distributions from contact-tracing data. *bioRxiv*, 683326.

Tian, S., N. Hu, J. Lou, K. Chen, X. Kang, Z. Xiang, H. Chen, D. Wang, N. Liu, D. Liu, et al. (2020). Characteristics of COVID-19 infection in Beijing. *Journal of Infection*.

Zhao, S., P. Cao, M. K. Chong, D. Gao, Y. Lou, J. Ran, K. Wang, W. Wang, L. Yang, D. He, et al. (2020). 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. *Infection Control & Hospital Epidemiology*, 1–8.

Zhao, S., D. Gao, Z. Zhuang, M. Chong, Y. Cai, J. Ran, P. Cao, K. Wang, Y. Lou, W. Wang, et al. (2020). 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. *medRxiv*.