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

Abstract

1 Introduction

In December 2019, a cluster of pneumonia cases of unknown etiology was reported in China. The disease, now referred to as the coronavirus disease 2019 (COVID-19), has since affected more than ??? countries. As of XXX, more than XXX deaths have been confirmed.

Skipping introduction for now.

Understanding the natural history of a disease is crucial for predicting the course of an outbreak and controlling it. At the population level, this is often done by summarizing the time between key epidemiological events as probability distributions. These events can be compared within an infected individual (e.g., infection and symptom onset) or between infected individuals (e.g., symptom onsets of an infector and an infectee). However, the estimation of these epidemiological delay distributions depends on having observed both events; this dependency can bias the estimates during an early growth phase of an outbreak.

2 Theoretical framework

In order to understand how the observed delay distributions between two epidemiological events vary over time, we begin by defining the epidemiological delays from a cohort 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 consists of all individuals who became symptomatic at the same time.

Cohort-based delay distributions are always subject to right-censoring. Since both events must occur before the time of measurement t to be observed, delays that are longer than $t - s$ cannot be observed for a cohort at time $s < t$. Therefore, 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 (which can vary across cohorts), $F_s(\tau)$ is the corresponding cumulative distribution function, and $0 \leq \tau \leq t-s$. While delay distributions measured within an individual (e.g., incubation period) may not vary much over the course of an epidemic, those measured between individuals vary due to changes in the susceptible population. Within-individual delay distributions that depend on external factors (e.g., time between symptom onset and hospitalization) can also change over time.

Typically, epidemiological delay distributions are estimated by using all available samples. Then, the observed delay distribution $f_{\text{obs}}(\tau|t)$, which takes into account all measured delays until time t, is equivalent to the 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:

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

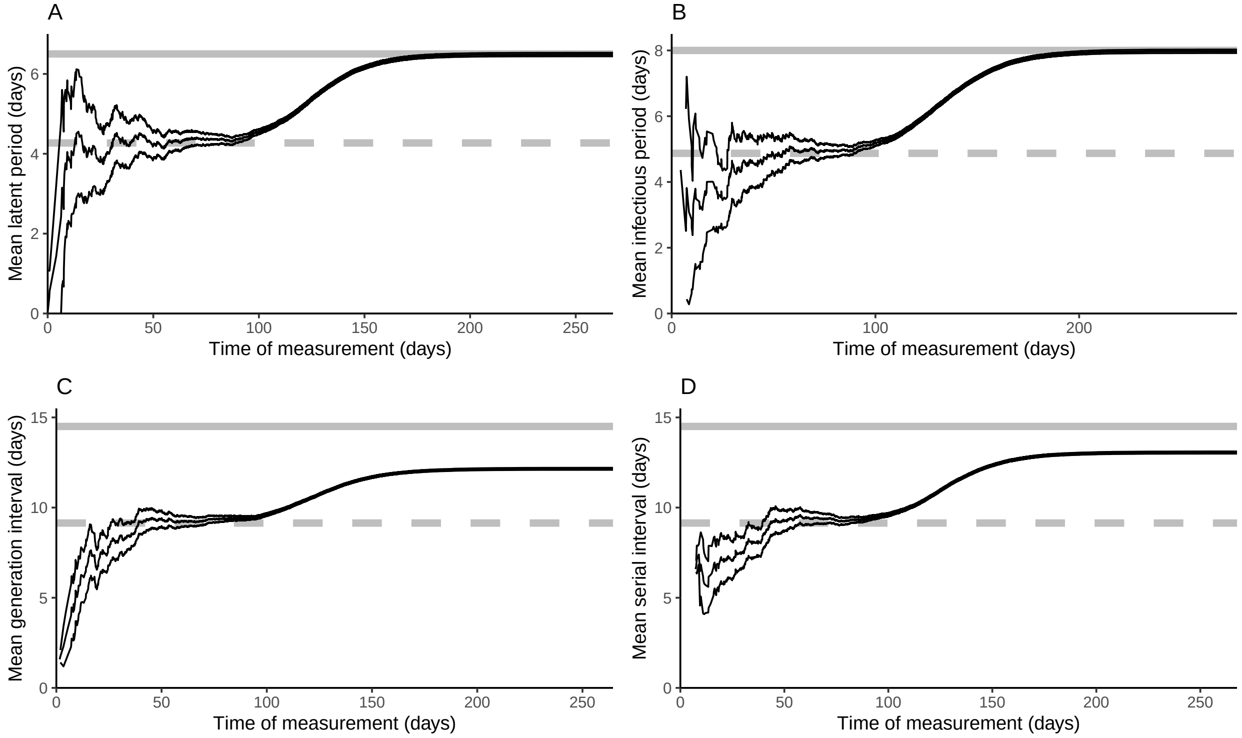

Figure 1: **Observed means of epidemiological delay distributions over time.** Changes in the observed mean latent period (A), infectious period (B), generation interval (C), and serial interval (D) over the course of an epidemic. Black solid lines represent the observed mean and associated 95% confidence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

Early in an epidemic, the incidence of infection, and therefore the incidence of cohorts, is expected to grow exponentially: $i(s) = i(0)\exp(rs)$. We can assume that the true delay distribution stays constant for both within- and between-individual delays during this period: $f_s(\tau) = f(\tau)$. Then, the observed delay distribution during the exponential growth phase $f_{\text{exp}}(\tau|t)$ is equivalent to the true distribution deweighted by the exponential growth rate:

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

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.

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)

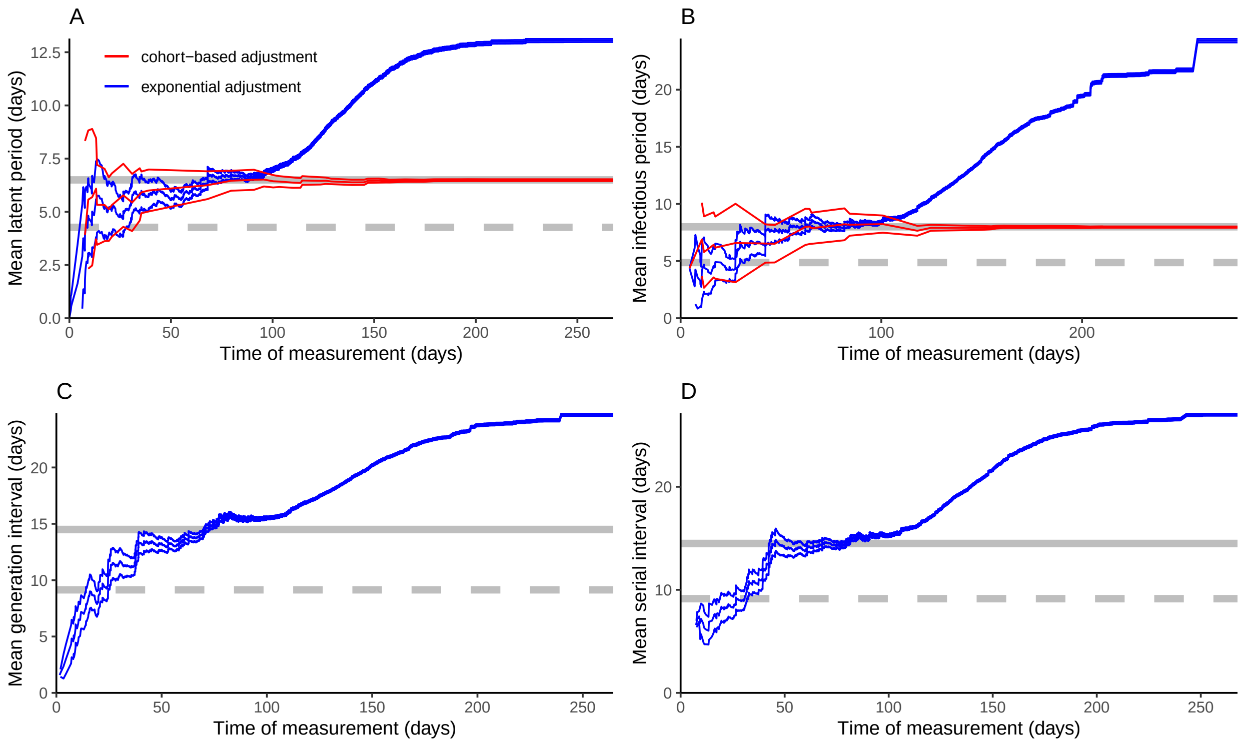

Figure 2: **Estimated means of epidemiological delay distributions over time.** Changes in the estimated mean latent period (A), infectious period (B), generation interval (C), and serial interval (D) over the course of an epidemic. Black solid lines represent the estimated mean and associated 95% confidence intervals at each time of measurement using the exponential adjustment. Red solid lines represent the estimated mean and associated 95% confidence intervals at each time of measurement using the cohort-based adjustment. Gray solid lines represent the theoretical true mean. Gray dashed lines represent the theoretical observed mean during the exponential growth phase (calculated via (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 $I(0) = 10$.

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 theoretical mean that we calculate using (4). The amount of bias decreases as the epidemic progresses; eventually, the observed mean latent and infectious periods become unbiased. On the other hand, mean mean observed generation and serial intervals remain biased even at the end of an epidemic because the depletion of the susceptible pool over time makes the realized generation intervals shorter over time. The serial interval has a slightly higher observed mean than the generation interval near the end of an epidemic because an infected individual with a short latent period (and hence shorter incubation period) is more likely infect others by having a shorter generation interval during the susceptible depletion phase; therefore, infectors are more likely to have shorter latent periods than their infectees.

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

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

Similar forms have been suggested by other studies and applied in estimating epidemiological delay distributions; however, our simulations show that this naive approach does not work very well for estimating the mean (Fig. 2; exponential adjustment). Although the estimated mean delay is similar to the true mean, the estimates are sensitive, and their associated confidence intervals are narrow. Once the exponential growth phase is over, applying the exponential adjustment overestimates the mean delay.

Alternatively, we can account for the bias by ensuring that the right-censoring does not exist: instead of using all samples that have been collected until the time of measurement, we can limit our samples to cohorts within which all individuals have completed both epidemiological events of interest. This approach gives an unbiased estimate of the mean delay throughout the epidemic with appropriately wide confidence intervals that contain the true value (Fig. 2; cohort-based adjustment). Conversely, the proportion of individuals that have not completed the second event within each cohort is 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 (i.e., we cannot measure the degree of right-censoring). Nonetheless, likelihood-based methods that explicitly account for the right-censoring can be still applied to estimate the generation interval distribution.

3 Applications: incubation period distribution of COVID-19

Here, we revisit the incubation period distribution estimated by Backer et al. (2020) and assess the degree of bias in their estimate. Since they relied on early traveller data who travelled from Wuhan between January 2–23, during which the epidemic was likely to have been expanding exponentially, we can expect the number of infected travellers from Wuhan to increase exponentially; therefore, their estimate of the mean *observed* incubation periods of 6.4 days (95% CI: 5.6–7.7 days) is likely to have been subject to right-censoring that we describe earlier.

4 Discussion

5 Materials and Methods

5.1 Stochastic simulations

See Park, Champredon, and Dushoff (2019) in bioRxiv.

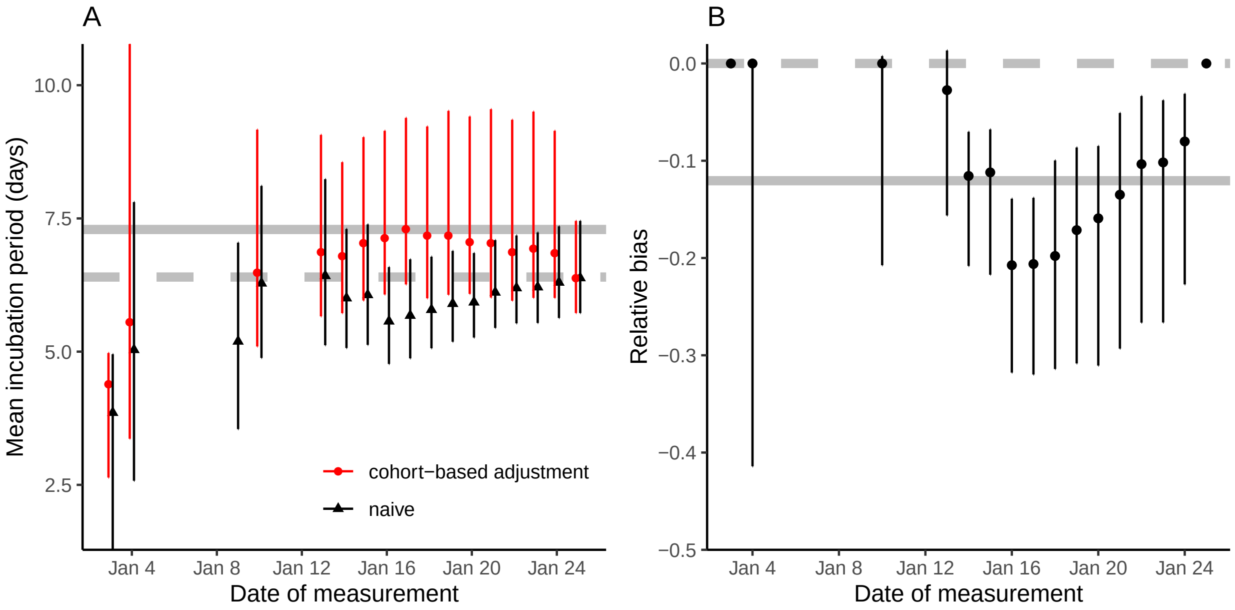

Figure 3: **caption.** caption

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).