# Towards robust and accurate estimates of the incubation time distribution, with focus on upper tail probabilities and SARS-CoV-2 infection

Arntzen, V.H.; Fiocco, M.; Leitzinger, N.; Geskus, R.B.

Received: 16 May 2022 | Revised: 17 February 2023 | Accepted: 18 March 2023

DOI: 10.1002/sim.9726

RESEARCH ARTICLE

Statistics in Medicine WILEY

# Towards robust and accurate estimates of the incubation time distribution, with focus on upper tail probabilities and SARS-CoV-2 infection

**Vera H. Arntzen**[1] | **Marta Fiocco**[1,2,3] | **Nils Leitzinger**[1] | **Ronald B. Geskus**[4,5,6]

[1]Mathematical Institute, Leiden University, Leiden, Netherlands

[2]Biomedical Data Science, Medical Statistics Section, Leiden University Medical Center, Leiden, Netherlands

[3]Trial Data Center, Princess Maxima Center for Childhood Oncology, Utrecht, Netherlands

[4]Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, UK

[5]Biostatistics, Oxford University Clinical Research Unit (OUCRU), Ho Chi Minh City, Vietnam

[6]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK

**Correspondence**
Vera H. Arntzen, Mathematical Institute, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, Netherlands.
Email: v.h.arntzen@math.leidenuniv.nl

**Abstract**

Quarantine length for individuals who have been at risk for infection with SARS-CoV-2 has been based on estimates of the incubation time distribution. The time of infection is often not known exactly, yielding data with an interval censored time origin. We give a detailed account of the data structure, likelihood formulation and assumptions usually made in the literature: (i) the risk of infection is assumed constant on the exposure window and (ii) the incubation time follows a specific parametric distribution. The impact of these assumptions remains unclear, especially for the right tail of the distribution which informs quarantine policy. We quantified bias in percentiles by means of simulation studies that mimic reality as close as possible. If assumption (i) is not correct, then median and upper percentiles are affected similarly, whereas misspecification of the parametric approach (ii) mainly affects upper percentiles. The latter may yield considerable bias. We suggest a semiparametric method that provides more robust estimates without the need of a parametric choice. Additionally, we used a simulation study to evaluate a method that has been suggested if all infection times are left censored. It assumes that the width of the interval from infection to latest possible exposure follows a uniform distribution. This assumption gave biased results in the exponential phase of an outbreak. Our application to open source data suggests that focus should be on the level of information in the observations, as expressed by the width of exposure windows, rather than the number of observations.

**KEYWORDS**

incubation time, interval censored data, quarantine period, SARS-CoV-2, semiparametric, uniform infection risk

## 1 | INTRODUCTION

Isolation of individuals with established SARS-CoV-2 infection and quarantining individuals with higher risk of infection (risk contacts) were two of the widely adopted policy measures to slow down the spread of the virus upon its emergence

in 2020. Quarantine length for risk contacts is commonly based on the incubation time distribution, that is, the time from infection to symptom onset.[1] At the beginning of the SARS-CoV-2 pandemic, the WHO recommended a quarantine length of 14 days from the last time that a risk contact was exposed to an infected individual.[2] Country-specific quarantine lengths deviate from the WHO recommendation, depending on the policy aim, such as 'flattening the curve' in the Netherlands and–until July 2021– 'zero spread' in Vietnam. A stricter policy requires a longer quarantine period.

Because of its relevance for policy makers, empirical estimates of the incubation time distribution were made soon after the start of the SARS-CoV-2 pandemic, mostly based on early data from Wuhan, China. Xin et al performed a systematic review and meta-analysis of published studies until September 25, 2020 that reported mean, median and/or $95^{th}$ percentile of the incubation time along with 95% confidence intervals (CIs).[3] Individual estimates of the $95^{th}$ percentile ranged from 3.2 to 18.3 days. The pooled estimates of the $95^{th}$ percentile were dependent on the chosen parametric distribution: 12.6 days (7 studies; 95% CI, 11.2–14.0) and 14.1 days (5 studies; 95% CI, 12.3–15.8) for estimates based on lognormal distribution and Weibull distribution, respectively. These estimates concern the 'wild' type; in comparison, new variants seem to have a shorter incubation time.[4]

While the event time (time of symptom onset) is generally observed, the big challenge is knowing the time origin (time of infection). We mostly only know the start and end of the exposure window, and often just the end. This makes the infection time interval censored or left censored. Most studies assumed a uniform distribution of infection risk within the exposure period. This has the advantage that the time scale can be reversed and traditional methods for interval and right censored time-to-event data can be used. In case all infection times are left censored, this approach cannot be used because all data would be right censored on the reversed time scale. For such data, another approach has been suggested which makes assumptions similar to those in renewal process theory.[5,6]

While the systematic review by Xin et al focused on the estimates, the aim of our study is to review the methods used to estimate the SARS-CoV-2 incubation time distribution. We give a detailed account of the data structure, the likelihood formulation and the assumptions commonly made in the literature (Section 2). By means of simulations we studied the robustness of these assumptions (Sections 3 and 4). The focus of this paper is on the impact of these assumptions on the estimates of the median and upper tail percentiles, as the quarantine length is based on these quantities. As an illustration, percentiles of the SARS-CoV-2 incubation time distribution are estimated using openly available data from the first months of the pandemic (Section 5).

## 2 | DATA COLLECTION AND METHODS

### 2.1 | Data on infection time

For most infectious diseases it is difficult to obtain information on time of infection. Estimates of the HIV incubation time distribution were mostly based on data from cohort studies, where participants were tested for the presence of HIV antibodies once every three to 6 months. Since antibodies can be detected within the first months after infection and the median incubation time to AIDS is about ten years, using the midpoint of the seroconversion window as infection time gives negligible bias.[7]

The situation with SARS-CoV-2 is very different. Time from infection to symptom onset varies from a few days to a few weeks. Many symptoms for COVID-19 are not very specific, may have another cause and many individuals remain asymptomatic. Since antibodies develop several weeks after infection, diagnosis of acute infection is based on the RT-PCR test for the presence of RNA. Information on infection time is obtained from four possible sources:

1. time of direct contact with one or more infected individuals
2. a time period during which an individual was at risk of infection, without having information on specific contacts
3. time of a first positive RT-PCR test
4. time of symptom onset.

Intensive tracing of all contacts of diagnosed individuals can provide a rich source of information if the incidence of infection is low. However, data quality is hampered by recall bias of the time of contact, the presence of several possible infectors, and the true source of infection may even be missed. Also, in infection clusters where the possible contacts are known, it may be unknown who infected whom.

If no specific contacts are known, there may be general information on the window of exposure. Infected but still undiagnosed individuals can end a period of infection risk if they are quarantined and have no further contact with others. As an example, in the first 1.5 years of the pandemic, the Vietnamese government quarantined all direct contacts (F1) of an infected case (F0) in an allocated facility with active monitoring, and the second line (F2) of contacts at home. The earliest estimates[8,9] of the SARS-CoV-2 incubation time were based on individuals who left Wuhan before they developed symptoms. Assuming that the virus was absent outside Wuhan at that time, departure from Wuhan ends the exposure window. These individuals had a minimum incubation time which is the time span between departure and disease onset. Most studies additionally included individuals who arrived in Wuhan during the first outbreak in January 2020. This defines their start of the infection risk period and gives a maximum incubation time. If for nobody the start of the exposure window is known, additional assumptions as discussed in Section 2.3 are required.

A positive test result only provides information on the infection time if the person tested positive before symptom onset. An earlier negative test does not provide any information because the person may already be infected at that time. Although symptom onset is the end point of the incubation time, it provides information on an individual's maximum incubation time if the start of his exposure window is known.

Data on infection and symptom onset have primarily been collected for public health purposes to contain further spread of the virus and monitor individuals with symptoms. They have not been collected in a rigorous scientific way. Many studies are based on data from government websites, which lack detailed information on the data collection process, on the choices made with respect to allocation of infection source and on the definition of symptom onset used.

## 2.2 | Likelihood and assumptions for interval censored infection times

We start by describing the approach used when some individuals have an interval censored exposure window. For individual $i$ ($i = 1, \dots, N$), let $E_{il}$ and $E_{ir}$ be the calendar times that denote the start and end of the exposure window respectively (Figure 1). $E_{il}$ may be missing or set at a value before the start of the outbreak. Let $S_i$ be the calendar time of symptom onset.

Denote by $g_i(\cdot|e_{il}, e_{ir})$ the density of the infection time, given the individual's exposure window. We allow $g_i$ to depend on the individual. Let $f(\cdot)$ and $F(\cdot)$ be the density and cumulative distribution function of the incubation time and $h(\cdot, \cdot)$ the density of the exposure interval.

We assume incubation (E to S) and infection (E) time distributions to be independent. The contribution to the likelihood from individual $i$ is

$$l(e_{il}, e_{ir}, s_i) = h(e_{il}, e_{ir}) \int_{e_{il}}^{e_{ir}} g_i(t|e_{il}, e_{ir}) f(s_i - t) dt. \tag{1}$$

Note that commonly $E_{il}$, $E_{ir}$ and $S_i$ are observed up to a day precise, but this discretization is not taken into account in the likelihood.

The infection time distribution can be defined at the population level or at the individual level. For the earliest studies on the SARS-CoV-2 incubation time, no individual contact data was used, while the pandemic was in its exponential

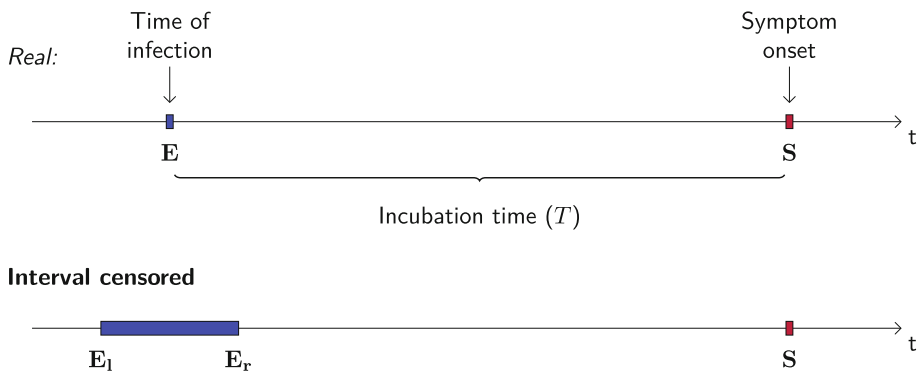


**FIGURE 1** Timeline for interval censored observations of incubation time (infection to symptom onset).

phase. This suggests using a single population wide distribution $g$, similar to what has been used for studies on the HIV incubation time where left, right and interval censored infection time are present.[10]

For SARS-CoV-2 there are several reasons to assume an individual exposure distribution $g_i(\cdot|(e_{il}, e_{ir}))$ instead. Infection rates can be very local and depend on the setting in which transmission occurred (at home, at work, in a public place). Also, contact rates fluctuate with an individual's willingness to comply with preventive and lockdown measures. And if contact tracing, precautionary quarantining and testing are related to suspected infection, then the time of infection is more likely to be close to the end of one's exposure window.

Most studies on the SARS-CoV-2 incubation time assume that the risk of infection is constant on the exposure window. Then, the contribution of $g_i(\cdot|e_{il}, e_{ir})$ to the likelihood reduces to a constant $(\frac{1}{e_{ir}-e_{il}})$ that can be left out, yielding

$$l(e_{il}, e_{ir}, s_i) \propto \int_{e_{il}}^{e_{ir}} f(s_i - t)dt = F(s_i - e_{il}) - F(s_i - e_{ir}). \tag{2}$$

and we end up maximizing

$$\sum_{i=1}^{N} log\left[F(s_i - e_{il}) - F(s_i - e_{ir})\right]. \tag{3}$$

Hence, by assuming a constant risk of infection on the exposure window, the time axis can be reversed and standard methodology for interval censored data can be applied. If the infection time is left censored, it is possible to treat it as interval censored by choosing the start of the exposure window far before the start of the outbreak. When the time axis is reversed, this transforms into a right censored observation.

The validity of assuming a constant infection risk can be questioned, as the risk within the exposure window may vary by calendar time (if the outbreak is in the exponential growth phase), location and by type of contact (eg, infection in a household or at work). We performed a simulation study to quantify the bias when the actual infection time is monotonically increasing or decreasing whereas a constant risk is assumed (Section 3).

Common practice is to describe the incubation time using parametric models such as lognormal, gamma and Weibull and choose the one that provides the most conservative tail percentile[1,11] or the best-fitting distribution based on AIC.[12] As there are few observations in the tail, the best fitting distribution and the parameter estimates will mainly be based on the larger amount of information in the middle of the distribution. As a consequence, the estimates of the tail percentiles will strongly depend on the assumed parametric distribution and the form of the incubation time distribution in the middle part. However, there is little biological evidence to support the assumption that the incubation time follows one single parametric distribution over the whole domain. A systematic review showed that the estimates of the $95^{th}$ percentile of the SARS-CoV-2 incubation time distribution vary according to the choice of the parametric model.[3] Also, confidence intervals for the tail percentiles will be too narrow if assumptions are made that are uncertain to hold. We investigated the presence of bias under the assumption of an incorrect parametric distribution in a simulation study (Section 3).

## 2.3 | Likelihood and assumptions for the approach of Qin, Deng et al

An analysis by Qin et al,[6] later refined by Deng et al,[5] used data from 1211 individuals who developed symptoms after they had left Wuhan during the first exponential phase of the outbreak. The authors only included individuals travelling between January 19 and 23 2020, which is the time from the first public awareness of the severity of the outbreak until the city's lockdown right before Chinese New Year. To further ensure that all individuals were infected in Wuhan, the authors excluded cases who left Wuhan with their infected relatives and friends.

For each subject $i \in \{1, \ldots, N\}$, date of travel $D_i$ and symptom onset $S_i$ were available (Figure 2). The incubation time $T_i$ can be written as the sum of the forward $V_i = S_i - D_i$ and backward time $A_i = D_i - E_i$. Since only the forward time $V_i$ is observed, additional assumptions need to be made to estimate the incubation time. The authors assumed that the time of leaving Wuhan can be seen as an observation time from a truncated renewal process that has reached the equilibrium state, where time of infection and symptom onset are renewal times. However, strictly speaking, this is not a renewal process, as there is no sequence of events of similar type. Rather, each individual has only two events of different type and their timelines overlap in calendar time.

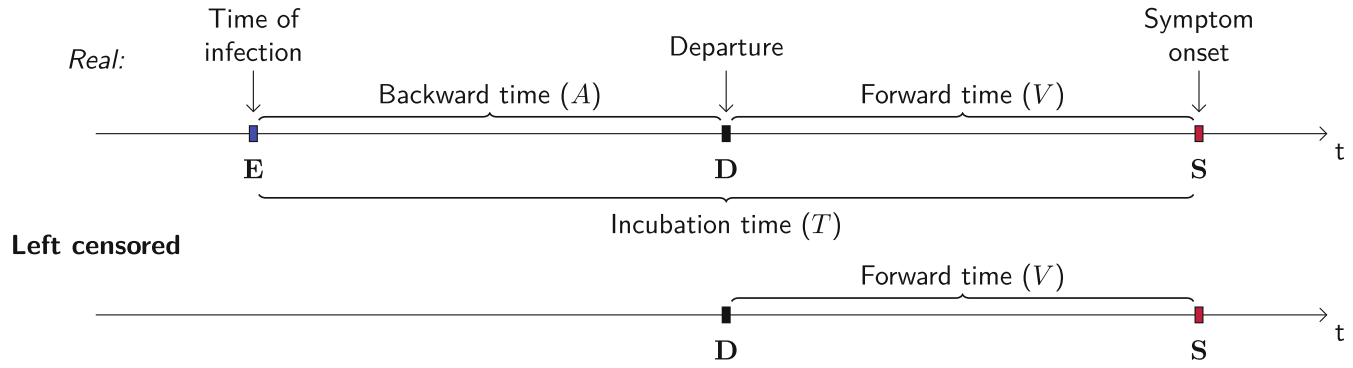


**FIGURE 2** Timeline for observations of forward time (departure from Wuhan, China, to symptom onset) to estimate incubation time (infection to symptom onset). As departure in this context marks the end of the exposure window and the start of the exposure window is unknown, the time of infection is left censored.

More precisely, they assumed that travel is independent of infection and symptom onset and occurred randomly after infection according to $A_i \sim U(0, \tau)$ with $\tau$ set at 30 days. Left truncation arose because individuals that developed symptoms while still in Wuhan were excluded from analysis. Denote by $f$, $F$ and $\mu$ the probability density function, cumulative distribution function and mean value of the incubation time distribution, respectively. Qin et al derived that the density of the forward time $h(v)$ for the individuals in the data set is

$$h(v) = \frac{1 - F(v)}{\mu} \quad 0 \leq v \leq \tau. \tag{4}$$

This implies that the density of the included forward time is monotonically decreasing. Since the backward time $A_i$ is assumed to follow a uniform distribution, it can be shown that the backward and forward time of the included data have the same distribution, denoted by $h(.)$.

The authors found that the forward times were not monotonically decreasing and therefore, they allowed for an additional risk of infection during travel, yielding the following mixture distribution for the observed forward times

$$q(v, \pi) = \pi f(v) + (1 - \pi)h(v), \quad v > 0 \tag{5}$$

where $\pi$ is the probability to get infected at the departure time from Wuhan. Note that the forward time $V_i$ equals the incubation time $T_i$ if infection occurs on the day of travel.

The estimation method proposed by Deng et al[5] improved Qin et al method[6] because it takes into account that the symptom onset day is essentially an interval of 24 h, due to the fact that the daily reports round information by day. Deng et al add and subtract 0.5 to each forward time, that is, $v_i^+ = v_i + 0.5$ and $v_i^- = v_i - 0.5$, yielding the following contribution of individual $i$ to the likelihood:

$$l(v_i) = \pi\{F(v_i^+) - F(v_i^-)\} + (1 - \pi)\{H(v_i^+) - H(v_i^-)\}. \tag{6}$$

In the remainder of this paper, only the method of Deng is discussed. Results from the method of Qin et al were very similar.

The validity of the assumption that travel occurs randomly between infection and day 30, that is, $A \sim U(0, 30)$, is uncertain. Since the outbreak started in a fully susceptible population without any prevention measures, the incidence was likely to increase exponentially. Also, many people left Wuhan in the few days before the lockdown and Chinese New Year. To investigate the robustness of the method in this particular context, we performed a simulation study with exponential growth of infection incidence and varying rates of leaving Wuhan (Section 4).

## 2.4 | Software

All analyses were performed in `R` version 4.1.1[13] and `R Studio` version 2021.09.20 ("Ghost Orchid")[14] software environment. `R` code is available from https://github.com/vharntzen/simstudy_incubationtime. This work was performed using

the computing resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University.

# 3 | SIMULATION STUDY I–INTERVAL CENSORED OBSERVATIONS

## 3.1 | Setup

Individual exposure window widths were sampled randomly from the observed exposure windows in five open source data sets[8,9,15,16] which we refer to as empirical widths.

To examine the impact of the assumption of constant risk of infection in the exposure window, three different infection risk distributions were simulated on the individual exposure windows: (i) constant risk ($g(t) \sim U(E_l, E_r)$), (ii) exponential growth with five-day doubling time of the incidence ($g(t) \propto e^{0.14t}$) which reflects the initial phase of the outbreak in Wuhan,[17] and (iii) a declining risk of transmission ($g(t) \propto p(1-p)^{t-1}$ where $p = 0.2$ on $[E_l, E_r]$), which may reflect household transmission. Figure 3B shows the risk functions for an exposure window of 10 days. The inverse cumulative distribution function (CDF) method was used. Moreover, to study how the impact of this assumption is affected by the exposure window width, more extreme scenarios were considered in which the widths were sampled after doubling or squaring the empirical widths.

To examine the impact of assuming an incorrect parametric distribution, incubation times were generated from a lognormal and Weibull distribution with parameters from Lauer et al.[9] We also generated incubation times from a more heavy-tailed Burr distribution, chosen such that the median was comparable to the two other distributions but with a considerably larger $95^{th}$ percentile (Figure 3C). Note that whereas the exposure window is discrete, no discretization was applied to the time of infection and symptom onset.

### 3.1.1 | Data generation

For each observation, an exposure window width ($E_r - E_l$), infection time $E$ and incubation time $T$ were sampled. Three different distributions of width were considered: the observed ("empirical") widths, all widths doubled and all widths squared (e.g. an observed width of 4 would become 8 and 16 respectively). Time of symptom onset $S$ was set to $S = E + T$. If $S < E_r$, we set $E_r = S$: COVID-19 related symptom onset determines the end of the exposure window as infection certainly took place before. For each scenario, 1000 data sets with size $N = 100$ and 500 were generated. Details about the algorithm can be found in Supplement A (Algorithm 1).

### 3.1.2 | Estimation

For each of the generated data sets, the $50^{th}$, $90^{th}$, $95^{th}$, $97.5^{th}$, and $99^{th}$ percentiles of the incubation time distribution were estimated using three parametric approaches, a semiparametric and a nonparametric approach. Maximum likelihood estimators (MLEs) assuming parametric distributions (gamma, lognormal and Weibull) were obtained by using the `flexsurv` package[18] while for the nonparametric maximum likelihood estimator (NPMLE) the `survival` package was used. For the semiparametric approach, a penalized Gaussian mixture (PGM) was employed using the `smoothSurv` package.[19] The smoothing factor $\lambda$ was chosen based on the maximum AIC in a sequence of values (0.1 to 5.6, or 0.5 to 5.5 with step size 0.5 for data sets of size 100 or 500, respectively) . More details are given in Supplement E.

For the percentiles in the parametric approaches, 95% CIs were obtained using parametric bootstrap as implemented in the `flexsurv` package. A method to obtain CIs for the NPMLE was explored (M out of N (M<N) bootstrap[20]). However, due to the size of the data set in this study, this technique was inadequate for the upper percentiles without smoothing, and CIs are not shown. Often, the estimate was equal to the upper limit of the CI. This is because the confidence interval disappears once the estimate of the cumulative distribution function reaches the value 1. More details can be found in Supplement D. For PGM, 95% CIs are obtained by basic bootstrap based on 1000 replications. To limit the computation time, instead of finding the optimal $\lambda$ for each bootstrapped data set, the $\lambda$ as obtained for the estimator itself was used for each bootstrap replication. In addition, for each single run (estimate including its bootstrapped confidence interval) we specified an upper time limit of three hours.

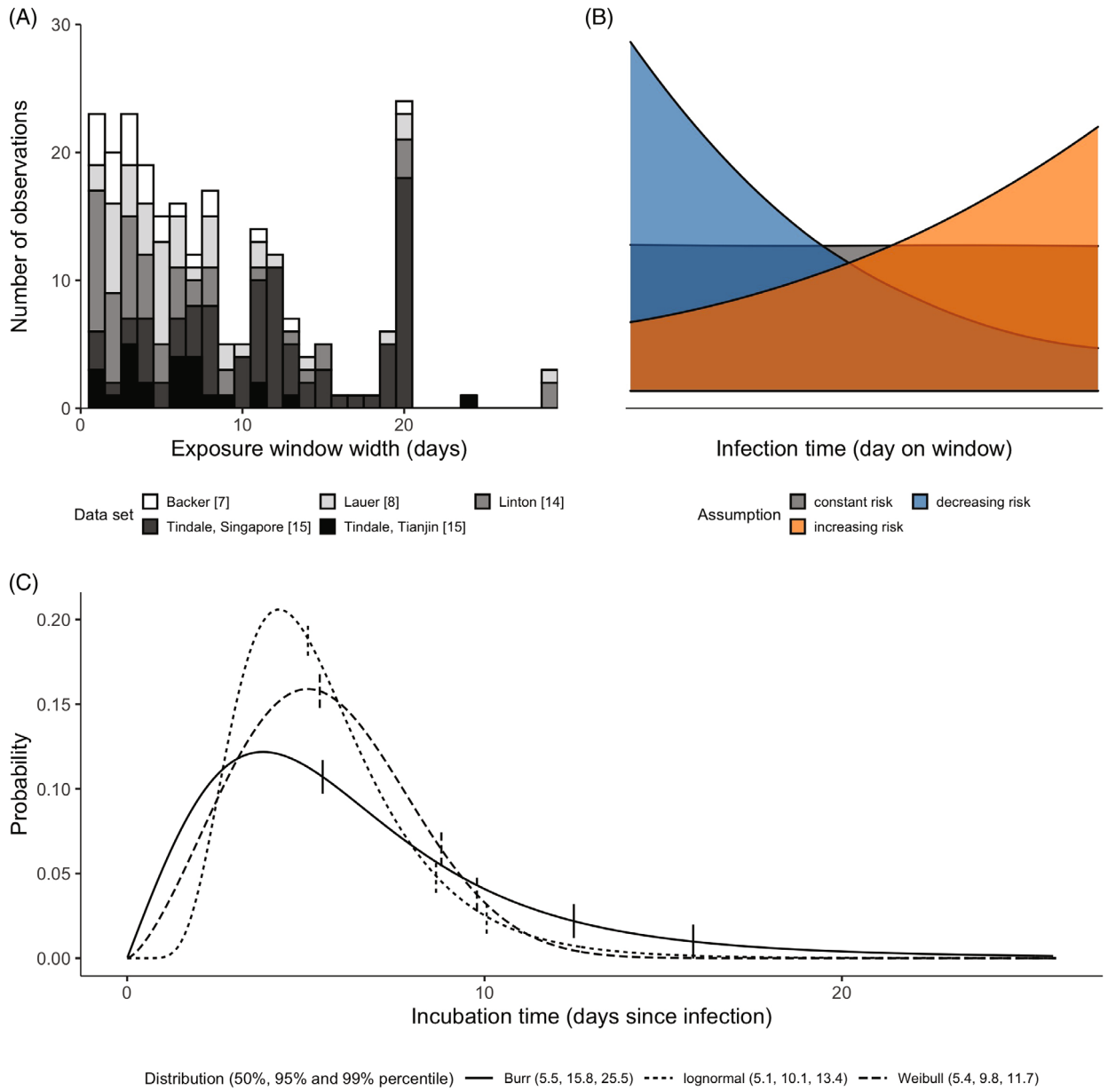


**FIGURE 3** Distributions used in simulation study I. (A) Distribution of exposure window widths in five openly available data sets from early in the pandemic. Shades of grey refer to the corresponding paper by author (and location).[8,9,15,16] (B) Infection risk distribution in the exposure window in three different scenarios, indicated by colour. (C) Distributions of incubation time used for simulation study and their median, 95$^{th}$ and 99$^{th}$ percentiles (represented by vertical bars). The parameterizations of Weibull (shape = 2.453, scale = 6.258) and lognormal (meanlog = 1.621, sdlog = 0.418) were based on SARS-CoV-2 specific estimates by Lauer et al;[9] the choice of Burr distribution is such, that its median is comparable to the other distributions, but the tail is more heavy (m = 8.5, s = 2, f = 2).

To assess the performance of the five estimation approaches, we report the mean deviation of the estimate from the true value (as estimate of bias), as well as the 25$^{th}$ and 75$^{th}$ percentiles of the deviations over all runs. We also report mean width of the 95% CI and its coverage.

## 3.2 | Results

In this section results are shown for sample size N = 100 and three percentiles. Results for N = 500 and other percentiles are provided in Supplement B.

## Incorrect parametric assumption introduces considerable bias in the tail estimates.

Figure 4 displays bias (A) and coverage probability (B) for the three chosen incubation time distributions and the five estimation approaches. The infection risk in the exposure window had a uniform distribution, as assumed in all five approaches. Estimates of the median and tail percentiles did not show any bias when the assumed distribution was the same as the true underlying distribution (see Figure 4A and Supplemental Figure B1A, middle and right panel, closed diamonds). Among all other combinations, NPMLE and PGM showed the smallest bias (cf. open diamonds). The variation (represented by vertical bars connecting quartiles of the estimates) in PGM was less than with the NPMLE, due to smoothing, and comparable to the parametric approaches. The incorrect parametric assumption led to a bias ranging from less than half a day for the median to more than three days in the $99^{th}$ percentile. The direction of the bias differed between the median and the tail percentiles. For example, for data generated from a Burr distribution (left panel), the bias was slightly upward for the median and downward for the tail percentiles when a gamma or Weibull distribution was assumed, and in the opposite directions when assuming a lognormal. As a consequence, there is a percentile between the median and the $99^{th}$ percentile where the estimate happens to be unbiased.

For most scenarios, the coverage deviated from 95% when the assumed distribution was different from the true one (Figure 4B and Supplemental Figure B1B, open diamonds). Coverage for the median dropped to lower levels when sample size increased (cf. Figure 4B and Supplemental Figure B1B, open diamonds). When the true distribution was Burr, different patterns were seen depending on the assumed distribution. Assuming a gamma or Weibull distribution yielded poor coverage when the percentile of interest was further towards the end of the tail. The lognormal distribution showed better coverage, especially in the tail for N = 100, where it was close to 95%. The latter may be related to the relatively wide confidence intervals (Supplemental Table B2). PGM showed fairly good coverage for all scenarios. For PGM, the coverage was based on a considerably smaller number of Monte Carlo replications but with at least 500 for each scenario (Supplemental Tables B1–B6, rightmost column).

## If the true infection risk in the exposure interval strongly deviates from uniform, assuming a constant risk of infection on the exposure window is only appropriate when intervals are relatively narrow.

Figure 5 shows (A) bias in estimates of median and tail percentiles and (B) coverage probability when the infection risk distribution on the individual's exposure window is monotonically increasing (exponential growth) or decreasing (household transmission), respectively. The first resulted in consistent overestimation (see Figure 5A and Supplemental Figure B2A, middle panel), whereas the latter showed consistent underestimation (cf. Figure 5A, right panel), when the incubation time distribution was either chosen correctly (diamonds, Weibull) or modeled flexibly (circles, PGM). Note that for some parametric distributions, the two different biases discussed in this paper cancel each other out (Supplemental Tables B1–B6). Violation of the uniform assumption led to similar bias in the median as compared to the tail percentiles, when the parametric assumption was correct (diamonds in Figure 5A, middle and right panel). This contrasts what was seen for the incorrect parametric assumption (cf. Figure 4A), namely that tail percentiles were more heavily affected than the median. Bias differed by exposure window width. Assuming a constant risk of infection in situations where this assumption is violated led to bias up to 3 days with a monotonically increasing infection risk on exposure windows of squared width (cf. Figure 5A, middle panel, right column).

Under a constant risk of infection (Figure 5A, left panel), even though the uniform assumption holds, the bias in the PGM estimate of the tail percentiles was around one day or larger. This is a consequence of the penalty that is imposed to guarantee smoothness of the distribution. For PGM, likewise parametric approaches, the tail behaviour is partly extrapolated from the part of the distribution where there are more observations, but for PGM this extrapolation is more local. Moreover, some of the bias may be due to the limitation discussed in Supplement E. However, in most scenarios this residual bias was smaller than with the incorrect parametric choice (Supplemental Tables B1–B6).

For both approaches, coverage deviated from 95% when risk of infection was not constant on the exposure window (see Figure 5B and Supplemental Figure B2B). Under exponential growth (middle panel), by using the empirical exposure window size the coverage was good assuming Weibull and poor using PGM, due to differences in bias. In case of declining risk (right panel), the coverage for Weibull was poor and for PGM was good. Coverage was considerably lower for the median than for the far right tail (cf. first vs. third rows) even though bias in the estimates using the correct distribution (represented by diamonds) is similar across percentiles. This is because the coverage proportion for the percentiles does

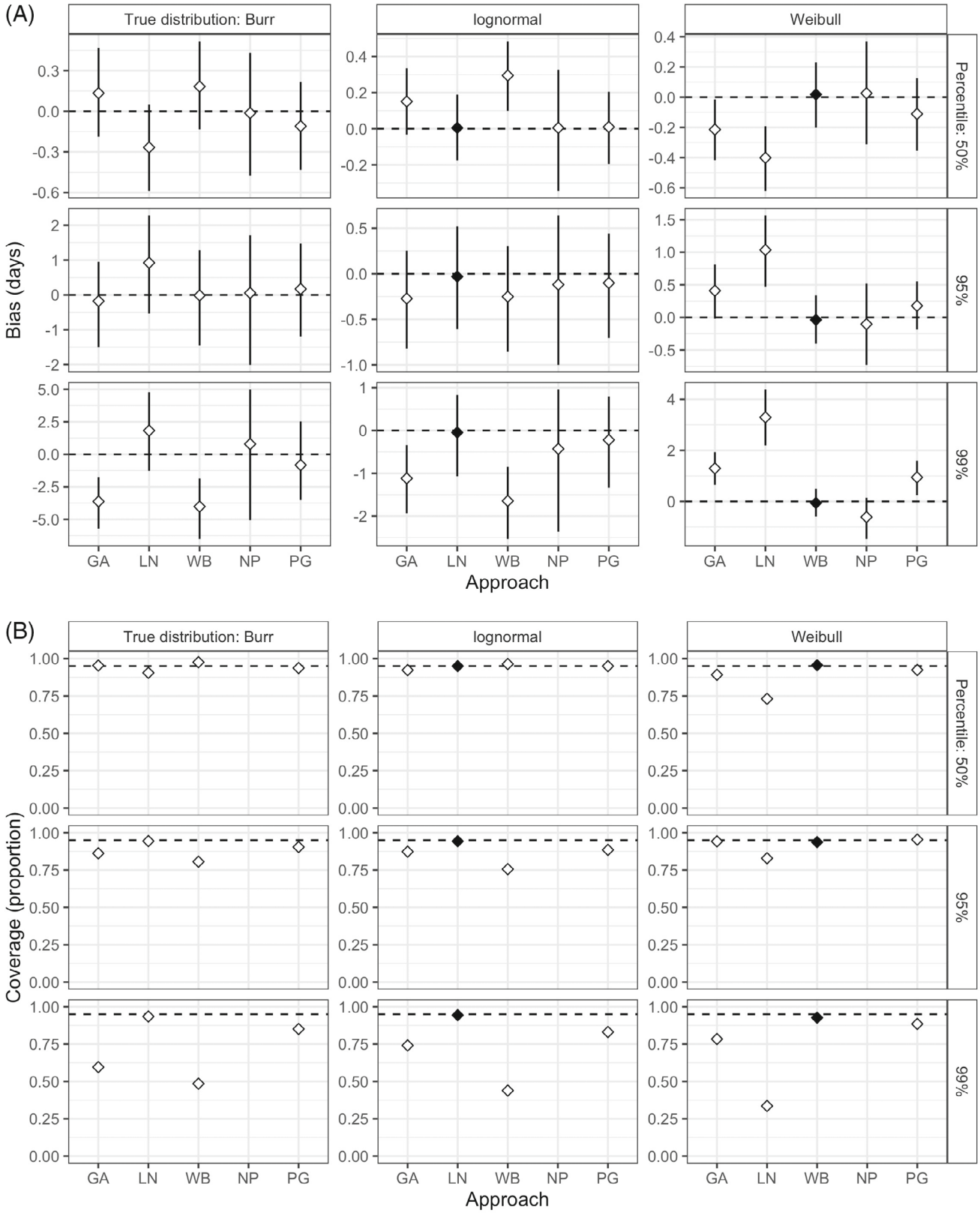


**FIGURE 4** Results of simulation study investigating the impact of assuming an incorrect parametric distribution of incubation time: bias (A) of estimated percentiles and (B) coverage of 95% confidence intervals. Vertical bars represent the inter quartile range of the deviation between estimate and true value. Five different estimation methods were used (x-axis): maximum likelihood estimator (MLE) assuming a gamma (GA), lognormal (LN) or Weibull (WB) distribution, nonparametric MLE (NP) and penalized Gaussian mixture (PG), respectively. Incubation times were generated from Burr, lognormal and Weibull distribution and a constant infection risk on the exposure window was assumed. Data set size: N = 100.

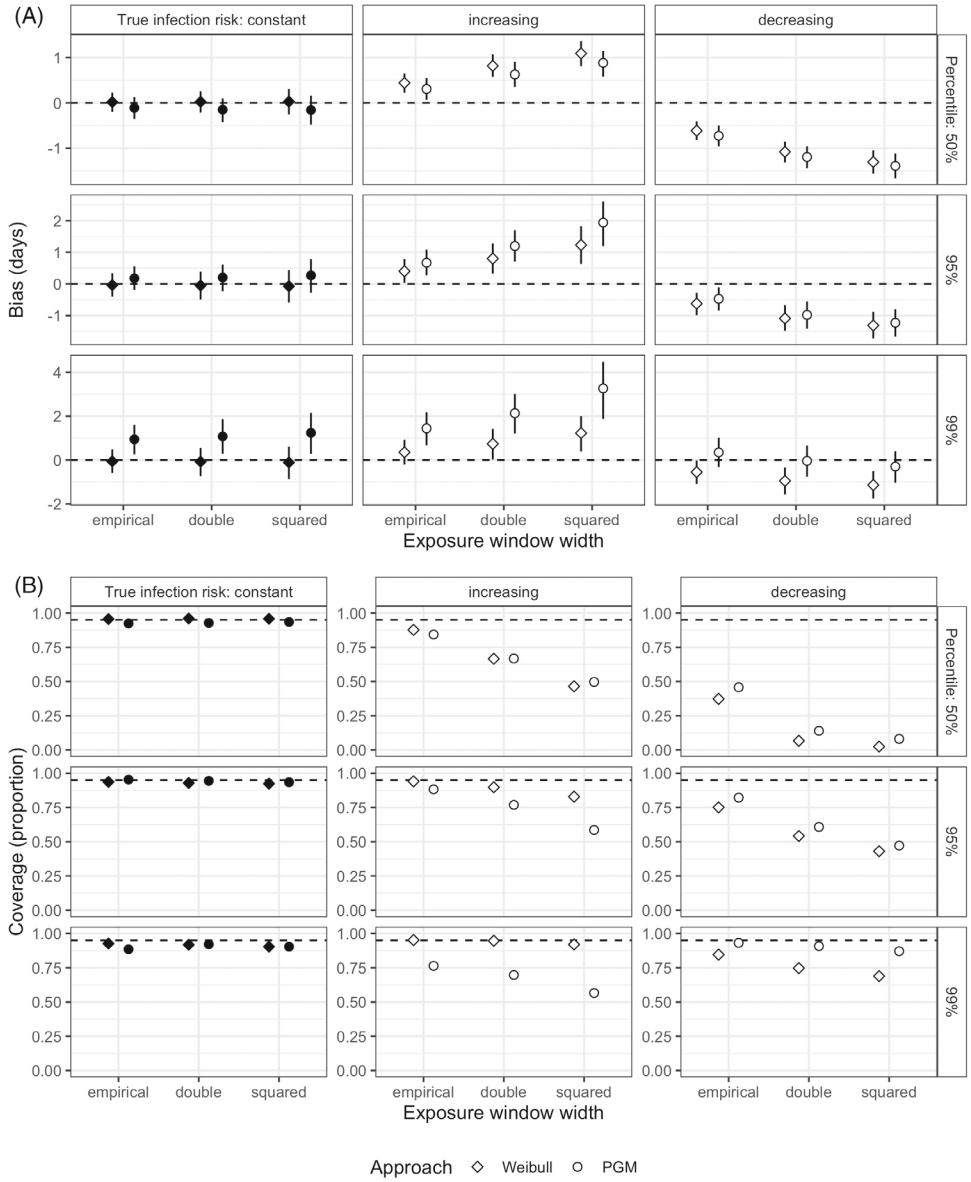


**FIGURE 5** Results of simulation study investigating the impact of assuming a constant risk of infection: (A) bias and (B) coverage proportion of percentiles (rows) estimated by MLE assuming Weibull and PGM model (shapes). Vertical bars represent the inter quartile range of the deviation between estimate and true value. Data was generated using different infection risk distributions (panels) and exposure window widths (x-axis). Incubation times were generated from the Weibull distribution. Data set size: N = 100.

not solely depend on the bias, but on the length of the confidence intervals as well. In fact, the CIs for the median were much smaller (Supplemental Tables B1–B6).

# 4 | SIMULATION STUDY II–RENEWAL PROCESS

## 4.1 | Setup

### 4.1.1 | Scenarios

In this simulation study, we investigated the validity of the approach of Deng,[5,6] and in particular the impact of the assumption that leaving Wuhan occurred randomly after infection. We simulated the early weeks of the outbreak in

Wuhan, assuming that the incidence of SARS-CoV-2 infections grew exponentially and that the rate of individuals leaving Wuhan sharply increased as the lockdown approached.[21] We also repeated the data generation process as in the simulation study by Deng,[5] in which no epidemic curve was assumed.

### 4.1.2 | Data generation

For each scenario, we chose sample sizes of N = 500 and N= 1200 and generated 1000 data sets. Incubation times were drawn from lognormal and Weibull distributions with parameters as estimated by Lauer et al[9] and a more heavy-tailed Burr distribution, chosen such that the median was comparable to the two other distributions but with a considerably larger $95^{th}$ percentile (Figure 3C).

`R` code for generating data according to the method proposed by Deng is available on https://github.com/naiiife/wuhan (accessed 06/30/2020). Travel days were drawn from a uniform distribution on domain [0,30], with infection as time origin 0. Denote by $\pi$ (with $\pi = 0, 0.1, 0.2$) the additional infection probability due to the travel. If an individual was infected before departure, then incubation time $T_i$ and travel time $c_i$ were drawn repeatedly until $T_i > c$, This implies that only individuals with symptom onset after travel were included. More details are provided in Algorithm 1.

**Algorithm 1.** Algorithm to generate observations of forward time $V$ of SARS-CoV-2, as proposed by Deng

**Result:** Data set with $i = 1, 2, \dots N$ observations of forward time $V$, where N = 500 or 1200.
**for** $i \leftarrow 1$ ***to*** $N$ **do**
  draw $D \sim Bernoulli(\pi)$;
  **if** $D = 1$ *(infected during travel)* **then**
    **return** $V_i \sim lognormal(\dots, \dots) *$ **or** $Weibull(\dots, \dots) *$ **or** $Burr(\dots, \dots, \dots)$;
  **else**
    **repeat**
      draw $T_i \sim lognormal(\dots, \dots) *$ **or** $T_i \sim Weibull(\dots, \dots) *$ **or** $T_i \sim Burr(\dots, \dots, \dots)$;
      draw $c_i \sim U(0, 30)$;
    **until** $T_i > c_i$;
    $V_i \leftarrow T_i - c_i$;
    **return** $V_i$;
**end**

Our alternative generation method (Algorithm 2) mimicked the infection and travel processes in the 18 days between January 5 and the lockdown of Wuhan on January 23, 2020. Resembling the population of Wuhan, we assumed 10 million susceptibles as initial population. As in the real data from the study of Deng, only those who travelled between January 19 and 23 and developed symptoms afterwards were included. We took January 5, 2020 as a starting date as those infected before were not likely to meet the criteria. Each day from January 5 to 18, the same number of people (150,0000) entered and left Wuhan. From January 19 to 23, no individuals entered Wuhan, but outbound travelling rate increased to 300,000 per day. The number of new infections on January 5 was chosen to be 125. The daily incidence of SARS-CoV-2 increased according to a five-day doubling time.[17] For the infecteds, incubation times were drawn and discretised using `R` function `round()`. This can be interpreted as all events (infection, symptom onset, travel) happening at noon. We selected individuals who left Wuhan between January 19 and 23, and developed symptoms during or after their day of travel. The travelling rates and initial number of new infections were chosen such that this yielded approximately 1200 observations, comparable to the real data used by Deng (1211 observations). Additionally, from each data set a smaller data set was obtained by randomly sampling 500 observations. Note that the probability of travelling was unrelated to infection status.

### 4.1.3 | Estimation

For each of the generated data sets, the $50^{th}$, $90^{th}$, $95^{th}$, $97.5^{th}$, and $99^{th}$ percentiles of the incubation time distribution were estimated using maximum likelihood estimation assuming gamma, lognormal and Weibull distributions.

**Algorithm 2.** Algorithm to generate observations of forward time $V$, taking into account the exponential growth of SARS-CoV-2 incidence and the sharp increase in people leaving Wuhan, China, before the lockdown. The rate of people entering and leaving Wuhan and initial number of new infections were chosen such that this yielded approximately 1200 observations

**Result:** Data set with $\approx$ 1200 observations of forward time $V$.
**1. Initialize**;
$P \leftarrow 1 : 10,000,000$ (population of Wuhan);
$S \leftarrow 1 : 10,000,000$ (susceptible population of Wuhan);
$E, D, V$ infection day, travel day and forward time of infecteds;
$I_0 \leftarrow 125$ (number of newly infecteds on January $5^{th}$);
**2. Infection and travel process**;
**for** $t \leftarrow 1$ ***to*** *19 (January $5^{th}$ to $23^{rd}$, 2020)* **do**
  *infect* $I_0 e^{0.14(t-1)}$ *with indices* $K \subset S$;
  $E[K] \leftarrow t$;
  $S \leftarrow S[-K]$;
  **if** $t$<15 *(before January $19^{th}$)* **then**
    *add 150,000 (people entering Wuhan)*;
    $S \leftarrow S[+150,000]$, $P \leftarrow P[+150,000]$ ;
    *remove 150,000 (people leaving Wuhan with indices* $M \subset P$*)*;
    $P \leftarrow P[-M]$;
    $S \leftarrow S[-(M \cap S)]$;
  **else if** $t \geq 15$ *(January $19^{th}$ to $23^{rd}$)* **then**
    *remove 300,000 people leaving Wuhan, with indices* $M \subset P$;
    $P \leftarrow P[-M]$;
    $S \leftarrow S[-(M \cap S)]$;
    $R_t \leftarrow$ indices $M \cap (P \setminus S)$;
    $D[M] \leftarrow t$;
    **for** $\forall i$ *in* $R_t$ **do**
      $T[i] \leftarrow$ *lognormal*$(\dots, \dots)$ **or** *Weibull*$(\dots, \dots)$ **or** *Burr*$(\dots, \dots, \dots)$;
      **if** $(E[i] + T[i]) \geq D[i]$ **select** *(i.e. left Wuhan between Jan $19^{th}$ and $23^{rd}$ and symptom onset on or after travel)*;
      $V[i] \leftarrow E[i] + T[i]) - D[i]$;
    **end**
**end**
**If** N = 500 **then** $V \leftarrow$ sample 500 from $V$;
**return** $(V)$

For each percentile, we report the mean deviation of the estimate from the true value (as estimate of bias), as well as the $25^{th}$ and $75^{th}$ percentiles of the deviations over all runs (Figure 7). For the mixture approach, we report the average estimate of $\pi$ and its 95% CI based on a normal approximation, conforming Deng. Coverage probabilities for the percentiles are not reported as our main interest is in the bias, bootstrapping as proposed by Deng is computationally demanding, and the authors did not provide the coverage in their work either.

## 4.2 | Results

In this section results from the estimation method of Deng and data set size N $\approx$ 1200 (range: 1064 to 1272) are discussed. Smaller sample size (N = 500) showed similar trends. See Supplement C for the additional results. With the assumptions as made by Deng and $\pi = 0$, the forward and backward time densities were equal and monotonically decreasing. This is no longer true when data were generated as in our new approach. See Figure 6.

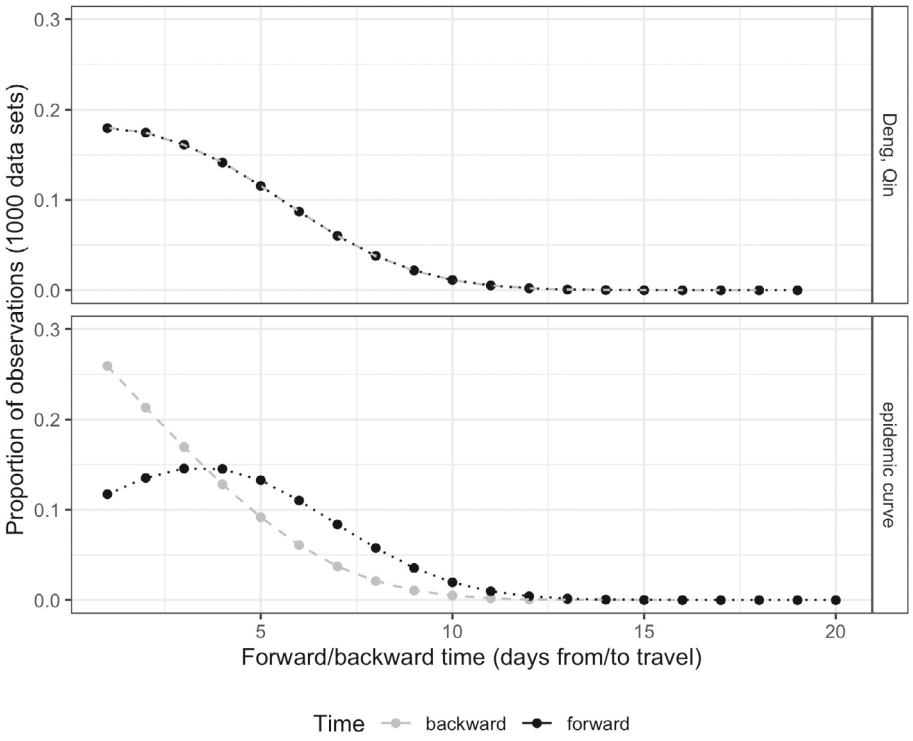


**FIGURE 6** Summary of forward and backward times (black and grey) in 1000 combined data sets generated by two different approaches (panels) for the following settings: sample size approximately 1200, $\pi = 0$, incubation time follows a Weibull distribution. Note that in the upper panel the two curves overlay.

## Little to no bias in estimates when travel day is sampled from uniform distribution and (mixture) model is correctly specified.

Results based on the data generation approach of Deng are shown in Figure 7. Figure 7A,C visualize the bias and mean of the estimates of $\pi$ in the mixture approach. Figure 7E shows the bias in the approach without mixture component. Colours refer to the distribution chosen to generate the incubation times given on the x-axis. If for data generated with $\pi = 0.2$ the correct parametric distribution was chosen, median, tail percentiles and $\pi$ show little to no bias (see Figure 7A,C, right panel, filled diamonds). Similarly, for the correctly chosen parametric distribution, the estimates didn't show any bias when data was generated with $\pi = 0$ and the model did not include a mixture component (cf. Figure 7E, left panel, filled diamonds). When data was generated with $\pi = 0$ and a mixture model was fitted little to no bias was observed either. However, when data was generated with $\pi = 0.2$ but analyzed without a mixture component, even the correctly chosen parametric model was strongly biased (Figure 7E, right panel).

## When data generation incorporates epidemic and travel trends, estimates of median and tail percentiles were heavily biased, even when the correct parametric distribution was chosen.

When no mixture component was included in the likelihood, the bias in the percentiles was considerable (cf. Figure 7F). Per contra, when data was generated with $\pi = 0$ and a mixture model was fitted, this bias was reduced (cf. Figure 7B, filled diamonds). Our alternative data generation process makes infection close to travel more likely whereas a uniform distribution is assumed in the model. Hence, the model gives an upward bias in the incubation time distribution. Allowing for additional infections on the day of travel via the mixture approach can capture some of this model misspecification. It even gives a downward bias because the true infection date is mostly before the day of travel. It yields large estimates of $\pi$ even though the data was generated without an excess risk while travelling (cf. Figure 7D).

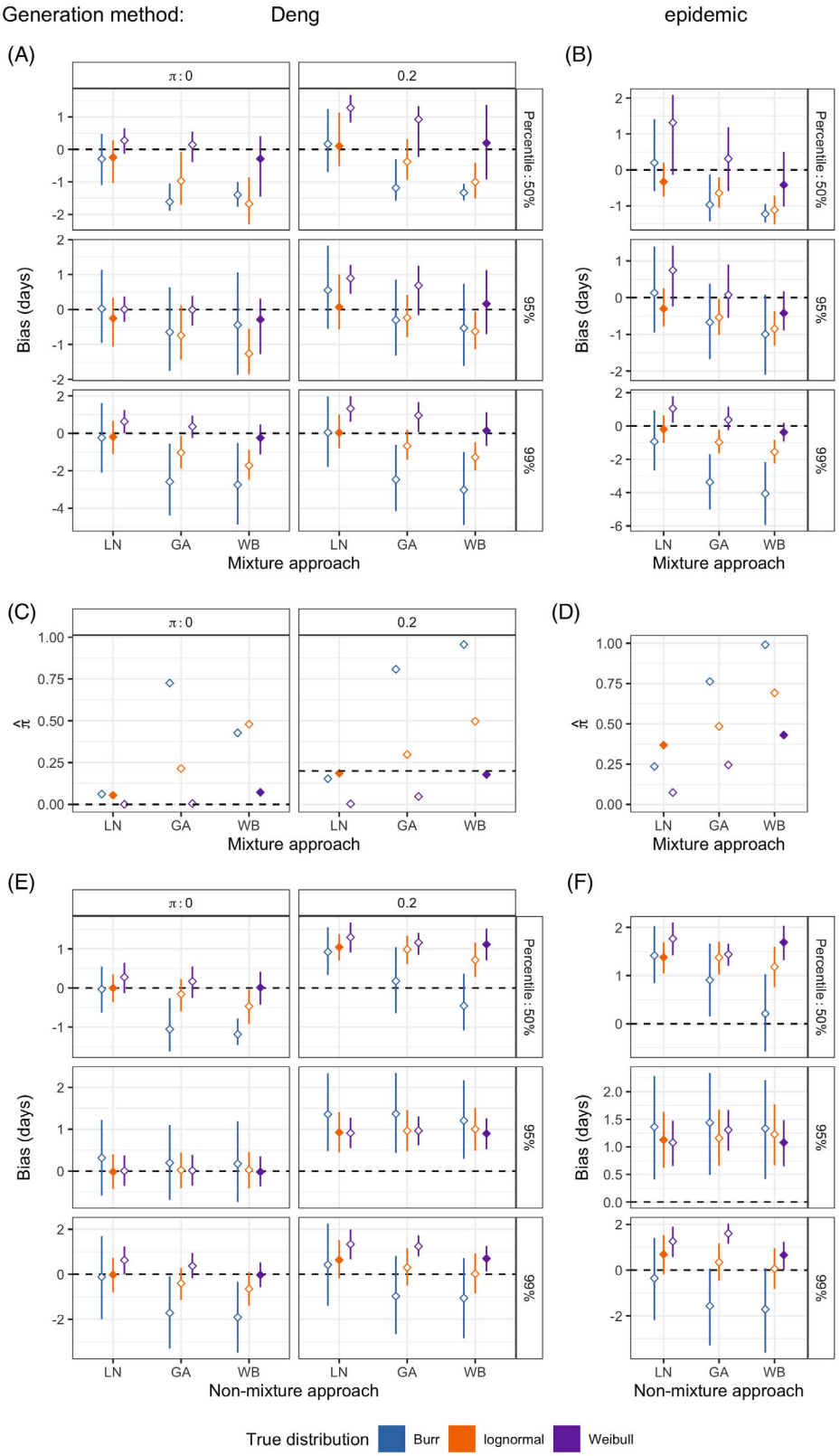


**FIGURE 7** Results from a simulation study investigating estimation of incubation time distribution using an method inspired by renewal process theory. Sample size is (approximately) 1200. Incubation times were generated from three different distributions: Burr (blue); lognormal (orange); Weibull (purple). Data generation: (A,C,E) Dengs method; (B,D,F) new method, that is, epidemic outbreak with $\pi = 0$. Estimation approach: mixture including $\pi$ (A–D) or excluding $\pi$ (E,F). A,B,E,F and C,D show the bias in the estimate of the percentiles and the average estimate of $\pi$, respectively. Vertical bars represent the inter quartile range of the deviation between estimate and true value. Dashed lines represent either zero bias, or the $\pi$ with which the data was generated.

When the choice of the parametric distribution is incorrect, bias in tail estimates can be as high as four days.

Figure 7A,B,E,F show that for almost all scenarios, bias in the percentiles was larger when the wrong parametric distribution was chosen (represented by open diamonds). In particular when the incubation times were generated from a Burr distribution, estimates of median and tail percentiles were strongly biased and showed large variability. The parameter $\pi$ was strongly overestimated in the mixture approach for incorrect parametric distributions (see Figure 7C,D).

## 5 | DATA ILLUSTRATION

### 5.1 | Open source data

Six publicly available data sets[8,9,15,16,22] with observations collected between 2020/01/31 and 2020/02/29 were combined. Five data sets consisted of individuals infected in China; one data set concerned individuals with local transmission in Singapore as well. The sample size ranged from 52 to 178. Fifteen individuals with interval censored time of symptom onset from one data set[9] were excluded from the analysis. Excluding another 70 asymptomatic individuals, 836 individuals were used. We divided the observations into five groups: exactly observed day of infection, interval censored day of infection with exposure window size smaller than or equal to the median width of 4 days, interval censored with wider exposure window, and left censored without information on the incubation time (end of exposure window is before, or coincides with symptom onset, respectively). Figure 8A visualizes all included observations and Figure 8B shows the frequency of each observation type per data set. Following common practice in these studies, missing information on the

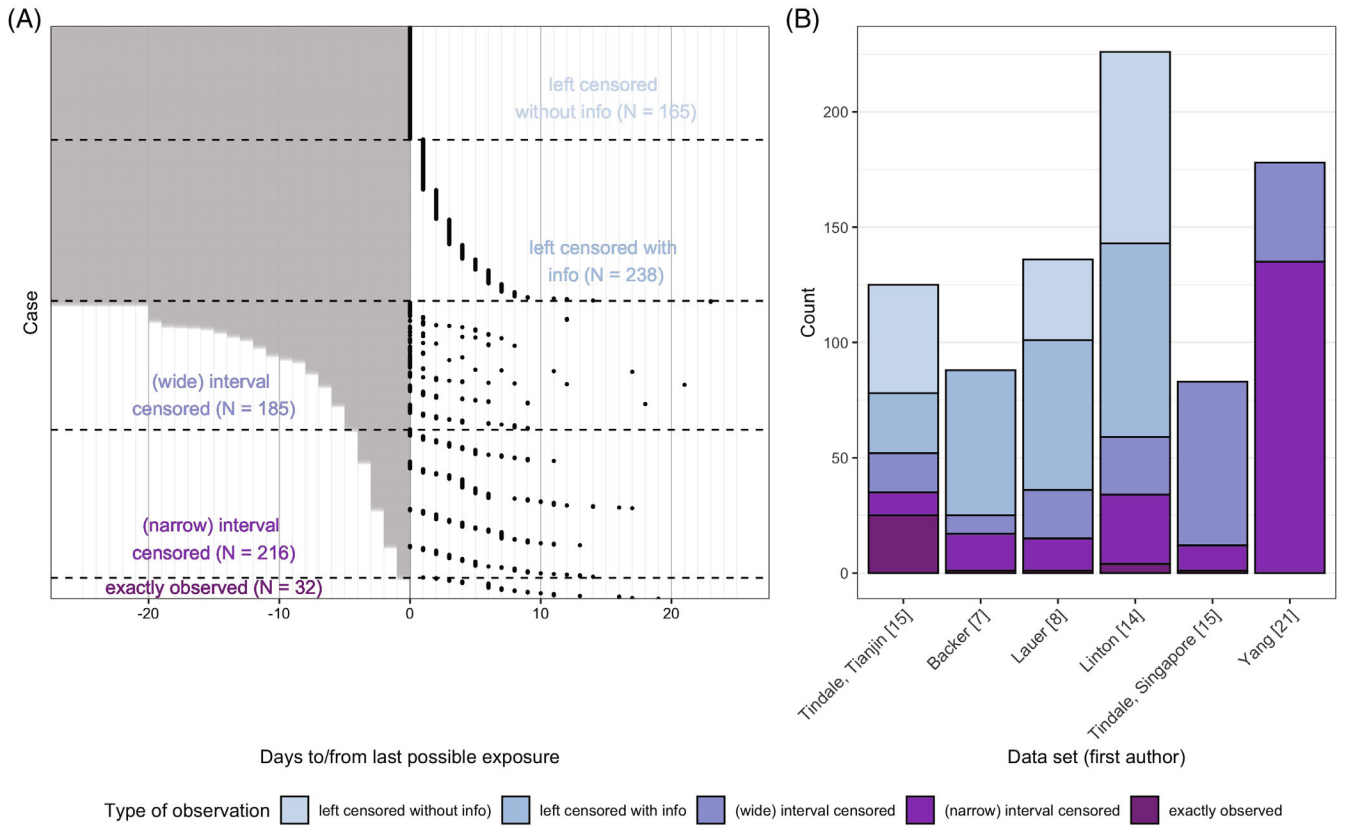


**FIGURE 8** Open source data to estimate the incubation time of SARS-CoV-2: (A) Visualization of individual timelines. Time is given as time from last possible exposure in days. Grey bars and dots indicate the exposure window and its midpoint (if any). Black dots indicate symptom onset. Cases are ordered by exposure window width and time between end of exposure and symptom onset. (B) Type of observations per data set. Data sets are indicated by first author. Bar height represents the number of observations in each data set.

start of exposure window (left censored observations) was replaced by December 31, 2019. Note that this is actually not needed, as the observations can be analyzed as left censored as well.

There are two important characteristics of the data that are beyond the scope of this paper but worth mentioning. First, individuals may have been included in multiple data sets. Second, as observations were collected while spread was ongoing, right truncation may occur: individuals who got infected shortly before the end of the follow-up of the study are only included if they have a short incubation period. This phenomenon leads to underestimation of quantiles of the incubation time distribution, that is stronger during an exponential growth phase. Linton et al[15] and Xin et al[3] accounted for right truncation, incorporating exponential growth in the number of infections over calendar time in the likelihood.

## 5.2 | Results

Figure 9 shows the estimates of the median and tail percentiles and their 95% CIs based on different partitions of the data, using the approach discussed in Section 2.2 (PGM was run with an AIC-based choice of $\lambda$ chosen from 0.5 to 5.5 with step size 0.5 and the same was done for the bootstrap runs to obtain 95% CI). For the NPMLE, confidence intervals were obtained as included in the function `survfit` from `R` package `survival`. We used the log-log transformation which made some of the upper bounds undefined. For more details, see Supplement D. The confidence intervals using only the 32 exact observations were quite wide and became much narrower after adding the 216 observations with narrow exposure windows as the amount of information increased. Estimates of the percentiles became smaller. This is likely due to the change in relative contributions of the data sets, which may have different characteristics, resulting in differences in corresponding estimates. The data set from Tianjin, China, by Tindale et al contains the majority of exact observations while the one by Yang et al contains most of the (narrow) interval censored observations. These authors estimated median

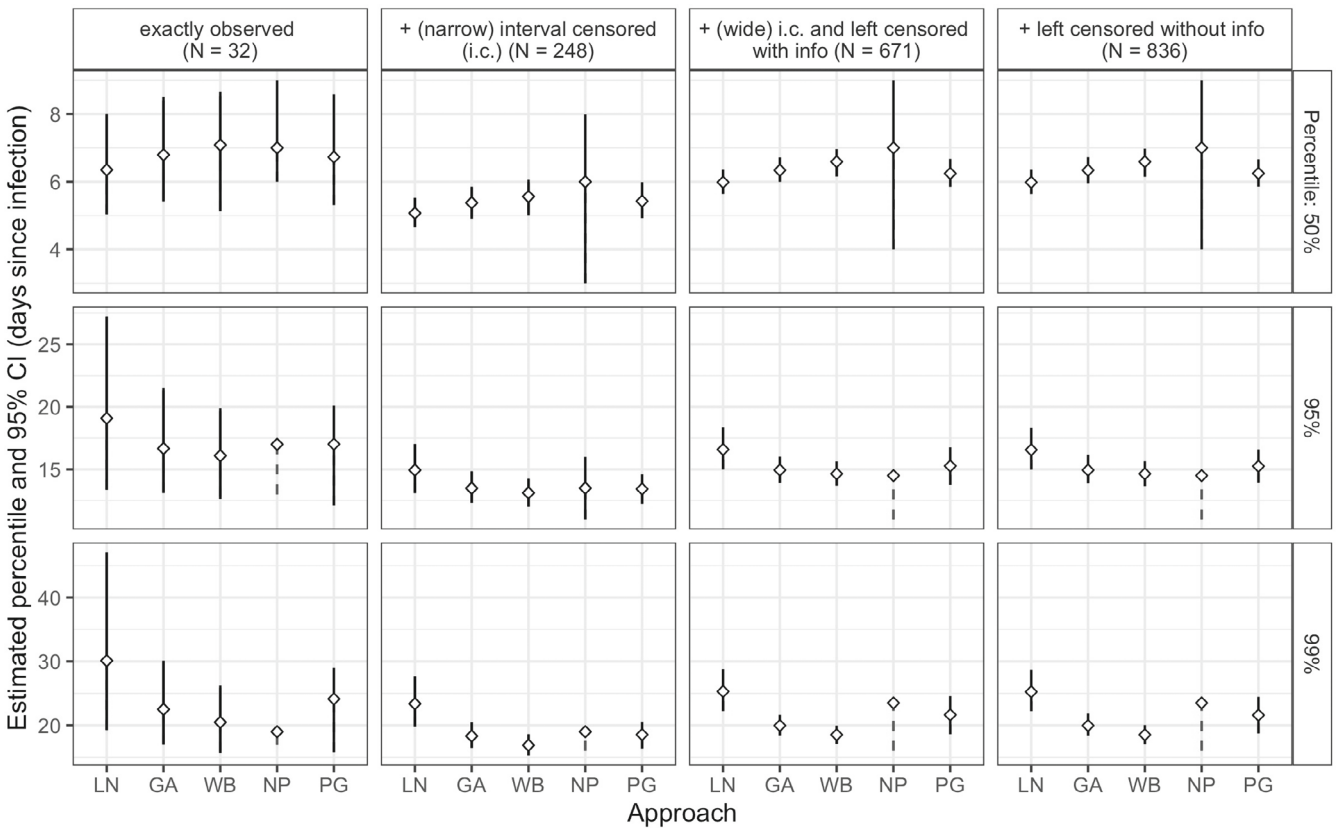


**FIGURE 9** Estimates of percentiles of the incubation time distribution based on different partitions of the data (panels) and estimation approaches (x-axis). Five different estimation methods were used: maximum likelihood estimator (MLE) assuming a gamma (GA), lognormal (LN) or Weibull (WB) distribution, nonparametric MLE (NP) and penalized Gaussian mixture (PG), respectively. Point estimates along with 95% CI are provided. When the upper bound of the CI was undefined, the lower bound is connected to the estimate instead (dashed line, NP).

incubation periods of 8.06 (95% CI 3.35; 5.72) and 5.4 (95% CI 4.8; 6.0), respectively. Next, we added the 185 individuals with wider exposure windows and 238 individuals for whom the start of exposure was unknown but the end of exposure was before symptom onset. This changed the estimates slightly in the upward direction, which may be explained by recall bias: more recent exposure (and infection) tends to be memorised better than exposure longer ago. Hence, individuals with a short incubation time are more likely to have a narrow exposure window than those with a longer incubation time. Therefore, estimates based on narrow exposure windows may be biased in the downward direction. The CI width didn't change much. Lastly, adding the 165 with only the end of exposure known changed the estimates and the CI widths very little. Note that they would not have changed at all if the infection times of individuals with unknown start of exposure window had been treated as left censored. This behaviour was seen in all estimated percentiles. For all methods, the width of the confidence intervals increased with the shift towards the (further) tail percentiles. Weibull and gamma approaches estimated the median higher than lognormal. For the tail percentiles this was reversed. Note that the estimate using the semiparametric PGM method was always in-between the two most extreme parametric estimates. This corresponds with our findings from simulation study I (Section 3).

## 6 | DISCUSSION

Estimates of the incubation time distribution have been and will be essential to inform policy makers at the start of an outbreak of a new pathogen like SARS-CoV-1 or SARS-CoV-2. The most challenging part is to obtain accurate data on the infection time. Most infection times are left or interval censored. We discussed and evaluated methods to estimate the incubation time of SARS-CoV-2. Our focus was on the commonly made assumptions and the resulting bias in estimates of the median and upper tail percentiles.

Most estimates are based on data sets in which not all infection times are left censored but at least some individuals have an interval censored infection time. To simplify estimation, standard practice has been to assume (i) a parametric distribution and (ii) a constant risk of infection within the exposure window. We examined the impact of both assumptions (i and ii) in a simulation study. Different parametric and nonparametric approaches (MLE assuming gamma, lognormal, Weibull; NPMLE) were considered. In addition, we proposed a semiparametric approach, that avoids the arbitrary choice of a parametric family yet preserves the smoothness of a parametric curve. We investigated the bias if the true infection risk in the exposure window is exponentially increasing (eg, an evolving outbreak) or declining (eg, household transmission). While an incorrect parametric choice mainly affected the upper percentiles, incorrectly assuming a constant risk affected the median and upper percentiles equally. We discuss the impact of each assumption in more detail.

Parameters are estimated based on all observations. The majority of observations is located in the middle. Accordingly, tail behaviour is forced to follow the behaviour in the middle of the distribution. For this reason, estimates of the tail percentiles were not robust to an incorrect parametric assumption of the incubation time distribution. In contrast to the pooled estimate of the $95^{th}$ percentile based on earlier estimates (13.1 days[3]), a recent study by Zhang et al[23] reported that more than 10% of individuals have an incubation time of more than 14 days. With our heavy-tailed Burr distribution, assuming a gamma or Weibull distribution estimated the $99^{th}$ percentile almost 4 days too small. The semiparametric approach proposed here–penalized Gaussian mixture- provides a good alternative. In smaller data sets it outperforms NPMLE, that often has the last jump to the value 1 before the upper percentiles of the true distribution. See Supplement D for details. However, the smoothing parameter $\lambda$ needs to be chosen carefully and the default procedure in the `smoothSurv` package did not always give satisfactory results (Supplement E). Moreover, for an incorrect parametric choice the confidence intervals tend to be too small for estimates of the tail percentiles. The confidence interval length for PGM and MLE assuming lognormal were fairly similar if the true distribution was Burr or lognormal (both heavy-tailed in our parameterisation). When the true distribution was Weibull, confidence interval length of PGM was most similar to the length obtained by the correct parametric choice.

The bias in the tail percentiles, introduced by falsely assuming a constant risk, tended to be smaller than the bias that can be attributed to the incorrect parametric choice. We saw that the bias increased with increasing average widths. If there is no recall bias, restricting to narrow exposure windows in the data gives the smallest bias, but it throws away information.

For many infectious diseases, like SARS-CoV-1 and Ebola, start of infectiousness coincides with or occurs after symptom onset. However, for SARS-CoV-2, 47.3% (95% CI: 34.0 to 61.0) of individuals remained asymptomatic throughout the course of infection,[24] while presymptomatic and asymptomatic transmission can occur.[16,25] Ideally, the distribution of time from infection to having detectable infection rather than incubation time should inform quarantine length for

potentially infected individuals. For many infectious diseases, this will be almost similar to the time from infection to start of infectiousness (latency time). The standard procedure to detect SARS-CoV-2 infection is to perform a PCR-test, giving rise to interval censored event times Estimation requires both a last negative and first positive PCR-test for at least part of the individuals. As both the start- and endpoint are interval censored (doubly interval censored data), estimation of these distributions is more complicated.[26] As a consequence, such estimates are rare and only became available later in the pandemic.[27]

Another way to avoid handling interval censored observations as such is restricting to exact observations or imputing or backtracing the infection moment. From literature it is known that restricting to exactly observed infection day leads to underestimation, due to recall bias: as individuals tend to recall more recent exposure more easily, observations of short incubation periods are more likely to be exact observations than those of long incubation periods. Midpoint imputation of the infection moment on the exposure window leads to bias as well.[1] Besides, it is only applicable when the start of the exposure window is known. A recent study[28] utilized viral load measurements to backtrace the actual infection moment but such data is usually lacking. Moreover, estimates may be biased if the strong assumptions that are made are incorrect.

Two earlier studies investigated the validity of assuming a parametric distribution and a constant risk on the exposure window.[1,26] Cowling et al noted discrepancies in the tails for different parametric models and named the nonparametric estimate as the gold standard.[1] In line with our results, Reich et al observed that for incorrect parametric choice, coverage for the median was lower as sample size increased.[26] The impact of assuming a constant risk on the exposure window was explored for a limited number of deviating risk distributions (piecewise uniform and spiked distribution) and two most extreme scenarios (all infected at beginning of the exposure window, all at the end).[1,26] The authors noted that the tail estimates were more sensitive to the choice of parametric distribution (i) than to the uniform assumption (ii), although performance was poor for the spiked distribution. This is similar to our findings. We contribute to their work by quantifying the bias for several scenarios inspired by SARS-CoV-2.

When data consist of only left censored observations of infection time, the above method based on the likelihood of interval censored data cannot be used. Alternative methods are needed, based on additional assumptions. Our second simulation study shows that the method proposed by Qin is not valid for the setting of an emerging outbreak.[6] Nevertheless, their method may be useful with data from individuals arriving in countries with a strict quarantine policy, if infection rates in the country of departure and arrival rates are relatively stable. Examples of such countries are Vietnam, China, New Zealand, Australia and Taiwan during part of the pandemic.

Qin et al performed a sensitivity analysis where they assumed an exponential density for the time from infection to departure. This led to an extra parameter in the likelihood of the forward times, which was estimated to be almost equal to zero. They concluded that the likelihood of forward times is approximately valid, even if the assumption of a uniform distribution of time from infection to departure does not hold. This is uncertain for two reasons. First, it was based on one real data set only. Second, it is difficult to assess how an exponential distribution for time from infection to departure relates to an exponential increase in the number of infections over calendar time which is much closer to what happened in reality. The authors compared their approach to the approach for interval censored data, but this comparison doesn't make sense. Their generation method for interval censored data leads to a different data set and it additionally has a conceptual mistake (see Supplemental Algorithm 2).

So far we discussed the marginal incubation time distribution, neglecting its dependence on covariables like age and comorbidities which may explain part of the differences between studies. Regression of incubation time on such covariables is needed to come up with a more personalized quarantine length, for example depending on age.[29] It is important to stress that also other factors are involved in choice of quarantine length. One example is the expectation of how people will adhere to it, which may be improved if quarantine period is shorter. Moreover, in countries that implement quarantining in allocated facilities, the capacity and economic costs may play a role. When policy makers aim for zero SARS-CoV-2 infections, more extreme percentiles than included in this paper might be needed to determine the length of the quarantine period. To accurately estimate percentiles in the far right tail, approaches based on extreme value theory may be more appropriate.

Quarantine length is based on the right tail of the incubation time distribution. Due to the interval censored time of infection, parametric assumptions are commonly made. We show that this can introduce only mild up to rather severe bias, mainly in the tail percentiles. Especially to inform quarantine length, a semiparametric method is a better option and can be used with available R Software. Whether the bias of the parametric methods is of clinical relevance depends on the aim of the policy. This cannot be seen separate from its societal context (ethics and resources) and disease characteristics (risk of severe outcome and transmissibility). Quarantine is a powerful measure for disease control, but cumbersome, and

requires accurate and congruent estimates in the literature to optimally inform decision makers. The penalized Gaussian mixture approach avoids the arbitrary choice between parametric distributions and reduces bias in tail percentiles.

The amount of smoothness can be chosen automatically using the standard procedure in the function `smoothSurvReg`.[30] If the resulting density shows a too wiggly pattern, the level of smoothness may be increased by increasing the value of the penalty term. We recommend to report the number of observations stratified by exposure window width, which gives additional information on the uncertainty in the estimate on top of the sample size.

## ACKNOWLEDGMENTS

The authors thank the COVID-19 modeling team at the Oxford University Clinical Research Unit, Vietnam: Duc Du Hong, Trang Duong Thuy, Lam Phung Khanh, Leigh Jones, Lieu Tran Thi Bich, Maia Rabaa, Manh Nguyen Duc, Marc Choisy, Nguyet Nguyen Thi Minh, Nhat Le Thanh Hoang, Sonia Lewycka, Thomas Kesteman, Trinh Dong Huu Khanh, Tung Trinh Son.

## DATA AVAILABILITY STATEMENT

R code and other sources that support the findings of this study are openly available in https://github.com/vharntzen/simstudy_incubationtime.

## ORCID

*Vera H. Arntzen* https://orcid.org/0000-0002-2642-9898
*Ronald B. Geskus* https://orcid.org/0000-0002-2740-3155

## REFERENCES

1. Cowling BJ, Muller MP, Wong IOL, et al. Alternative methods of estimating an incubation distribution: examples from severe acute respiratory syndrome. *Epidemiol (Cambridge, Mass.)*. 2007;18:253-259. doi:10.1097/01.ede.0000254660.07942.fb
2. World Health Organization. Consensus document on the epidemiology of severe acute respiratory syndrome (SARS), May 2003. 2003.
3. Xin H, Wong Jessica Y, Caitriona M, et al. The incubation period distribution of coronavirus disease 2019 (COVID-19): a systematic review and meta-analysis. *Clin Infect Dis*. 2021;73:2344-2352. doi:10.1093/cid/ciab501
4. Wu Y, Liangyu K, Zirui G, Jue L, Min L, Wannian L. Incubation period of COVID-19 caused by unique SARS-CoV-2 strains. *JAMA Netw Open*. 2022;5(8):e2228008. doi:10.1001/jamanetworkopen.2022.28008
5. Deng Y, Chong Y, Yukun L, Jing Q, Xiao-Hua Z. Estimation of incubation period and generation time based on observed length-biased epidemic cohort with censoring for COVID-19 outbreak in China. *Biometrics*. 2020;77:929-941. doi:10.1111/biom.13325
6. Jing Q, Chong Y, Lin Qiushi H, Shicheng TY, Xiao-Hua Z. Estimation of incubation period distribution of COVID-19 using disease onset forward time: A novel cross-sectional and forward follow-up study. *Sci Adv*. 2020;6(33):eabc1202. doi:10.1126/sciadv.abc1202
7. Gordon LC, Ron B. Effects of mid-point imputation on the analysis of doubly censored data. *Stat Med*. 1992;11(12):1569-1578. doi:10.1002/sim.4780111204
8. Backer Jantien A, Don K, Jacco W. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance*. 2020;25(5):1-6. doi:10.2807/1560-7917.es.2020.25.5.2000062
9. Lauer SA, Grantz KH, Qifang B, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann Internal Med*. 2020;172(9):577-582. doi:10.7326/m20-0504
10. Geskus RB. Methods for estimating the AIDS incubation time distribution when date of seroconversion is censored. *Stat Med*. 2001;20(5):795-812. doi:10.1002/sim.700
11. Nishiura H. Early efforts in modeling the incubation period of infectious diseases with an acute course of illness. *Emerg Themes Epidemiol*. 2007;4(2):1-12. doi:10.1186/1742-7622-4-2
12. Held L, Hens N, O'Neill P, Wallinga J. *Handbook of Infectious Disease Data Analysis*. London, UK: Chapman & Hall/CRC; 2019.
13. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna Austria: R Foundation for Statistical Computing; 2021.
14. RStudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, PBC; 2021.
15. Linton N, Kobayashi T, Yang Y, et al. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *J Clin Med*. 2020;9(2):538. doi:10.3390/jcm9020538
16. Tindale Lauren C, Stockdale Jessica E, Michelle C, et al. Evidence for transmission of COVID-19 prior to symptom onset. *eLife*. 2020;9:e57149. doi:10.7554/elife.57149
17. Dorigatti I, Okell L, Cori A, et al. Report 4: Severity of 2019-novel coronavirus (nCoV). 2020. doi:10.25561/77154
18. Jackson C. Flexsurv: a platform for parametric survival modeling in R. *J Stat Softw*. 2016;70(8):1-33. doi:10.18637/jss.v070.i08
19. Komárek A, Lesaffre E, Hilton JF. Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *J Comput Graph Stat*. 2005;14(3):726-745. doi:10.1198/106186005.63734
20. Lee SMS, Pun MC. On m out of n bootstrapping for nonstandard m-estimation with nuisance parameters. *J Am Stat Assoc*. 2006;101(475):1185-1197. doi:10.1198/016214506000000014

21. Gibbs H, Liu Y, Pearson CAB, et al. Changing travel patterns in China during the early stages of the COVID-19 pandemic. *Nat Commun.* 2020;11(1):5012. doi:10.1038/s41467-020-18783-0
22. 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. doi:10.1017/s0950268820001338
23. Zhang Z-J, Che T-L, Wang T, et al. Epidemiological features of COVID-19 patients with prolonged incubation period and its implications for controlling the epidemics in China. *BMC Public Health.* 2021;21(1):2239. doi:10.1186/s12889-021-12337-9
24. Pratha S, Fitzpatrick Meagan C, Zimmer Charlotte F, et al. Asymptomatic SARS-CoV-2 infection: a systematic review and meta-analysis. *Proc Natl Acad Sci.* 2021;118(34):e2109229118. doi:10.1073/pnas.2109229118
25. Van Vinh CN, Thanh LV, Thanh DN, et al. The natural history and transmission potential of asymptomatic severe acute respiratory syndrome coronavirus 2 infection. *Clin Infect Dis.* 2020;71(10):2679-2687. doi:10.1093/cid/ciaa711
26. Reich NG, Justin L, Cummings DAT, Brookmeyer R. Estimating incubation period distributions with coarse data. *Stat Med.* 2009;28(22):2769-2784. doi:10.1002/sim.3659
27. Xin H, Yu L, Wu P, et al. Estimating the latent period of coronavirus disease 2019 (COVID-19). *Clin Infect Dis.* 2021;74(9):1678-1681. doi:10.1093/cid/ciab746
28. Keisuke E, Kim Kwang S, Christina L, et al. Estimation of the incubation period of COVID-19 using viral load data. *Epidemics.* 2021;35:100454. doi:10.1016/j.epidem.2021.100454
29. Daewoo P, Klaus L, Jing N, Jordi Cortés Martínez, Gómez MG, Yu S. Modeling the coronavirus disease 2019 incubation period: impact on quarantine policy. *Mathematics.* 2020;8(9):1631. doi:10.3390/math8091631
30. Komarek A. smoothSurv package. 2020.


## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Arntzen VH, Fiocco M, Leitzinger N, Geskus RB. Towards robust and accurate estimates of the incubation time distribution, with focus on upper tail probabilities and SARS-CoV-2 infection. *Statistics in Medicine*. 2023;42(14):2341-2360. doi: 10.1002/sim.9726