Supplementary Text S1

1. Estimation of the time interval distribution using doubly interval-censored likelihood

Here we consider the fit of the incubation period as the time interval between the windows of exposure *“e”* and symptom onset *“s”* (written in lower or upper case). This approach can be applied to other time intervals, such as the interval between illness onset and death, illness onset and hospitalization, etc. The doubly interval-censored likelihood holds the following form:

|  |  |  |
| --- | --- | --- |
|  |  | (S1) |

where *L* refers to the left-hand (lower) value of a window of exposure or symptom onset, and *R* is the right-hand (upper) value; *i* is the index over all available data record of pairs and with the assumption that at least . A previously published work by Reich and colleagues [1] considered the same problem but under the condition of a uniformly distributed exposure distribution and . The results of Reich and colleagues can be altered to include a more general form of the function under the integral. Here, we describe how this can be done.

The main difficulty in calculating the likelihood (S1) is addressing the double integral over the rectangular area :

|  |  |  |
| --- | --- | --- |
|  |  | (S2) |

We notice that the integral can be transformed to the repeated integral after changing the variable , and subsequently calculated using Figure 1.

Figure 1A shows the situation when . In this case:

|  |  |  |
| --- | --- | --- |
|  |  | (S3) |

where is a cumulative distribution function (CDF) of .

Figure 1B shows the situation when . In this case:

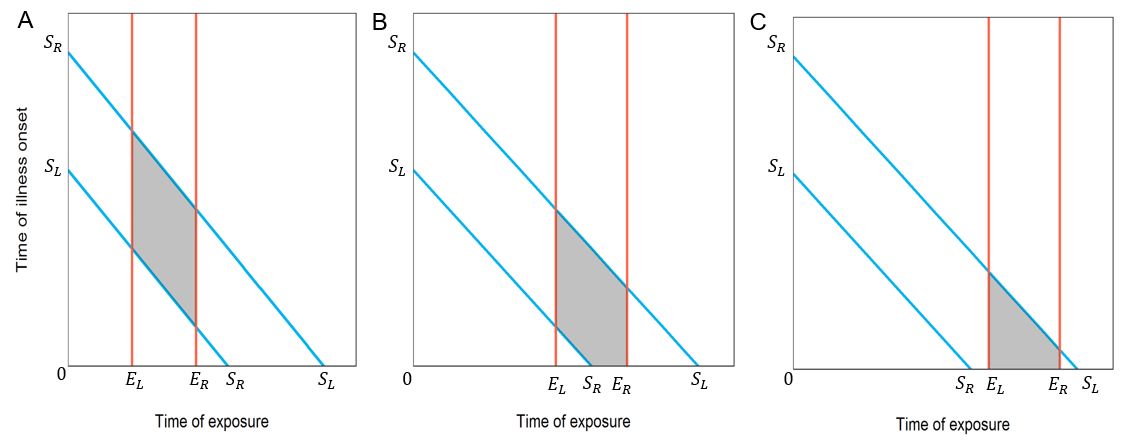
|  |  |  |
| --- | --- | --- |
|  |  | (S4) |

Lastly, Figure 1C shows the situation when In this instance:

|  |  |  |
| --- | --- | --- |
|  |  | (S5) |

We note that the doubly interval-censored likelihood with truncation admits the function to be separable on its arguments: . This means that the CDF: contains the CDF of the function and can be analytically derived for all three gamma, lognormal, and Weibull distributions considered in the main text.

Our numerical simulations confirmed that calculation of the likelihood using the formulas (S3)–(S5) lead to results that are identical to those produced using the R package coarseDataTools developed by Reich and colleagues [2].



**Figure S1. Illustration of the time interval distribution using doubly interval-censored likelihood.** The figure depicts how doubly interval-censored likelihood was calculated when A) , B) and C) . The blue lines and red lines indicate the lower or upper bounds of illness onset and exposure, respectively. The gray shaded area represents the integral in each situation.

# **B. Estimation of the time interval distributions using Bayesian framework.**

Here, we estimate parameters of time interval distributions for non-truncated likelihood using Markov chain Monte Carlo (MCMC) simulations. We used Stan packages with implementation of No U-Turn Sampler (NUTS). Specifically, each distribution is specified by the following prior distributions:

* The lognormal distribution is defined by the logarithms of the mean and the logarithm of the standard deviation. Both follow the standard normal distribution (with mean set at zero, and standard deviation equal to one).
* The Weibull distribution is defined by the logarithm of the mean and the logarithm of the shape parameter. Both follow the standard normal distribution.
* The gamma distribution is defined by the shape and inverse scale parameter. The former follows the normal distribution with the mean three and standard deviation five and is constrained to positive values. The latter follows a Cauchy distribution with location parameter set to zero and scale parameter equal to 5.0.

We followed general recommendations of the Stan developer community for the choice of prior distributions [3].

Additionally, we specified the priors for the times of exposure and illness onset (if necessary, relative to the other time interval distributions) for each case :

where: if and otherwise, and and are a pair of standardized variables:

on the interval between zero and one.

The non-truncated loglikelihood written in terms of the Stan language was defined as:

where is either the function , , or ; “lpdf” is shorthand for the logarithm of the probability distribution function.

We also used a slight variation of this model for the incubation period as to include censored likelihood to account for unknown lower bound of the exposure period.

In each case, we ran four chains of MCMC simulations consisting of 5,000 iterations for the simulation phase and 10,000 used for the tuning phase. The convergence of the obtained traces was verified visually, and we also explored the r-hat values and effective sample size. For example, Figure 2 shows the traces obtained for the fit of the incubation period to a lognormal distribution, which was the best-fit model as seen in Table 1 of the main text.



**Figure S2. The density plots and traces for the mean and standard deviation of the lognormal distribution when fitted to the dataset of incubation periods including Wuhan residents (second column of Table 1).**

1. Data cleaning rules implemented for the various time intervals

Incubation period:

For incubation periods of non-Wuhan residents (exposure type !=“Lives/works/studies in Wuhan”):

- Only include cases with travel to Wuhan, contact with Wuhan residents, or contact with a case.

- Must have a left exposure date (EL).

- One of onset, date first sought healthcare, or date hospitalized/isolated must be larger than left exposure date.

- For calculating ER and SL, must have right exposure date or onset date.

- For calculating SR, must have onset date or date first sought healthcare or date hospitalized/isolated or date of report.

- If left exposure (EL) is the same as the right exposure (ER)—which was the case for some people with day trips to Wuhan or contact with cases—we set ER to EL + 1 to provide a “day” of exposure.

- If right exposure date was not available or was greater than date of onset, ER was set to date of onset.

- Left onset (SL) is the date of onset (if available), otherwise set to the day after ER.

- Right onset (SR) is the day after date of onset if available, otherwise set to the date first sought healthcare, date hospitalized/isolated, or date reported as case, in that order.

For incubation period including Wuhan residents, the only difference in methods is that we do not require left exposure date for Wuhan residents. If left exposure date is NA, they must have exposure type “Lives-works-studies in Wuhan” or are excluded. Then, if the left exposure date is missing and they are a Wuhan resident, E\_L is set to 1 December 2019, which was the first date of onset for a reported case from Wuhan [4].

Onset to hospitalization:

- Calculated separately for deceased cases (almost all Hubei residents) and cases diagnosed outside of Hubei Province:

- Must have date of onset.

- Must have date of hospitalization or date of report (death for death dataset).

- Left onset (EL) is the date of onset.

- Right onset (ER) is the date of onset + 1.

- Left hospitalization/isolation (SL) is date of hospitalization, if available, otherwise date first sought healthcare, otherwise date of onset + 1.

- Right hospitalization/isolation (SR) is date of hospitalization + 1 if available, otherwise date of report (for export cases) or death (for deceased cases).

Onset to death:

- Must have date of onset.

- Must have date of death.

- Left onset (EL) is the date of onset.

- Right onset (ER) is the date of onset + 1.

- If time of death is known, left date of death (SL) is the date of death + time of death - 1 hour, otherwise date of death - 1 day.

- If time of death is known, right date of death (SR) is the date of death + time of death, otherwise date of death.

Hospitalization to death:

- Must have date of hospitalization or onset.

- Must have date of death.

- Left hospitalization/isolation (SL) is date of hospitalization, if available, otherwise date first sought healthcare, otherwise date of onset + 1.

- Right hospitalization/isolation (SR) is date of hospitalization + 1 if available, otherwise date of death - 1.

- If time of death is known, left date of death (SL) is the date of death + time of death - 1 hour, otherwise date of death - 1 day.

- If time of death is known, right date of death (SR) is the date of death + time of death, otherwise date of death.

# **References**

1. Reich NG.; Lessler J.; Cummings D.A.T.; Brookmeyer R. Estimating incubation period distributions with coarse data. Stat Med. 2009, 28:2769–84.
2. Reich NG.; Lessler J.; Azman A. coarseDataTools: A collection of functions to help with analysis of coarsely observed data. R package version 0.6-5, <https://cran.r-project.org/package=coarseDataTools>. 2019.
3. Stan developer team. Prior choice recommendations. <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>. Assessed 6 February 2020.
4. Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., … Feng, Z. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New England Journal of Medicine. 2020, 1–9. https://doi.org/10.1056/NEJMoa2001316.