A new framework and software to estimate time-varying reproduction numbers during epidemics: web material

Contents

Web Appendix 1. Estimation of the instantaneous reproduction number

Following Fraser , we assume that the distribution of infectiousness through time after infection is independent of calendar time. We model transmission with a Poisson process, so that the rate at which someone infected in time step $t-s$ generates new infections in time step *t*, is equal to $R_t w_s$, where R_t is the instantaneous reproduction number at time t and w_s a probability distribution (hence summing to 1) describing the average infectiousness profile after infection. Therefore the incidence at time t is Poisson distributed with mean $R_t \sum_{s=1}^{t} I_{t-s} w_s$, and the likelihood of the incidence I_t given the reproduction number R_t, conditional on the previous incidences $I_0,\ldots I_{t-1}$, is:

$$P\left(I_t \mid I_0,\ldots,I_{t-1},w,R_t\right)=\frac{\left(R_t\Lambda_t\right)^{I_t} e^{-R_t\Lambda_t}}{I_t!}$$

with $\Lambda_t=\sum_{s=1}^{t} I_{t-s} w_s$.

Now, if transmissibility is assumed constant over a time period $\left[t-\tau+1;t\right]$, measured by the reproduction number noted $R_{t,\tau}$, the likelihood of the incidence during this time period, $I_{t-\tau+1},\ldots,I_t$ given the reproduction number $R_{t,\tau}$, conditional on the previous incidences $I_0,\ldots I_{t-\tau}$, is:

$$P\left(I_{t-\tau+1},\ldots,I_t \mid I_0,\ldots,I_{t-\tau},w,R_{t,\tau}\right)=\prod_{s=t-\tau+1}^{t}\frac{\left(R_{t,\tau}\Lambda_s\right)^{I_s} e^{-R_{t,\tau}\Lambda_s}}{I_s!}.$$

Using a Bayesian framework with a Gamma distributed prior with parameters (a,b) for $R_{t,\tau}$, the posterior joint distribution of $R_{t,\tau}$ is

$$\begin{aligned}P\left(I_{t-\tau+1},\ldots,I_t,R_{t,\tau} \mid I_0,\ldots,I_{t-\tau},w\right) &= P\left(I_{t-\tau+1},\ldots,I_t \mid I_0,\ldots,I_{t-\tau},w,R_{t,\tau}\right)P\left(R_{t,\tau}\right)\\ &=\left(\prod_{s=t-\tau+1}^{t}\frac{\left(R_{t,\tau}\Lambda_s\right)^{I_s} e^{-R_{t,\tau}\Lambda_s}}{I_s!}\right)\left(\frac{R_{t,\tau}^{a-1} e^{-R_{t,\tau}/b}}{\Gamma(a)b^a}\right)\\ &=R_{t,\tau}^{a+\sum_{s=t-\tau+1}^{t} I_s-1} e^{-R_{t,\tau}\left(\sum_{s=t-\tau+1}^{t}\Lambda_s+\frac{1}{b}\right)}\prod_{s=t-\tau+1}^{t}\frac{\Lambda_s^{I_s}}{I_s!}\frac{1}{\Gamma(a)b^a}\end{aligned}$$

Which is proportional to:

$$R_{t,\tau}^{\,a+\sum_{s=t-\tau+1}^{t} I_s - 1} e^{-R_{t,\tau}\left(\sum_{s=t-\tau+1}^{t} \Lambda_s + \frac{1}{b}\right)} \prod_{s=t-\tau+1}^{t} \frac{\Lambda_s^{\,I_s}}{I_s!}.$$

Therefore, the posterior distribution of $R_{t,\tau}$ is a Gamma distribution with parameters $\left(a+\sum_{s=t-\tau+1}^{t} I_s, \frac{1}{\frac{1}{b}+\sum_{s=t-\tau+1}^{t}\Lambda_s}\right)$. In particular, the posterior mean of $R_{t,\tau}$ is $\frac{a+\sum_{s=t-\tau+1}^{t} I_s}{\frac{1}{b}+\sum_{s=t-\tau+1}^{t}\Lambda_s}$, and the posterior coefficient of variation (CV, standard deviation divided by mean) of $R_{t,\tau}$ is $\frac{1}{\sqrt{a+\sum_{s=t-\tau+1}^{t} I_s}}$.

The results shown in the main text were obtained using a Gamma prior distribution with mean 5 and standard deviation 5 (therefore $a=1, b=5$) for each $R_{t,\tau}$.

Web Appendix 2. Choice of time window

The estimates of R are expected to depend on the choice of the time window size τ. Small values of τ lead to more rapid detection of changes in transmission but also more statistical noise; large values lead to more smoothing, and reductions in statistical noise. So how to choose the appropriate time window?

Having an analytical formulation of the posterior distribution of R allowed us to link the posterior CV to the number of incident cases in the time window considered (see section Web Appendix 1). Imposing a posterior CV smaller than a predetermined threshold value $CV_{threshold}$ leads to $\sum_{s=t-\tau+1}^{t} I_s \geq \frac{1}{CV_{threshold}^{\,2}} - a$. This gives a minimum bound to the number of incident cases in the time window considered. Note that this result is independent on the infectiousness profile.

Web Table 1 presents the minimum number of incident cases in each time window corresponding to different choices of prior CV and aimed posterior CV. This result can provide guidance on the time windows to consider.

Web Table 1: Minimum number of incident cases in each time window as a function of the aimed posterior coefficient of variation (CV) for different choices of the prior coefficient of variation.

Prior CV	**Aimed posterior CV**									
	1	**0.9**	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0.3**	**0.2**	**0.1**
1	0	1	1	2	2	3	6	11	24	99
2	1	1	2	2	3	4	6	11	25	100
5	1	2	2	3	3	4	7	12	25	100
10	1	2	2	3	3	4	7	12	25	100
10000	1	2	2	3	3	4	7	12	25	100

Web Appendix 3. When can we start estimating *R*?

For real-time estimates of R to be useful, we need to start computing them early during an epidemic. However, estimating R too early in the epidemic might not be possible, at least if a certain precision in the estimates is desired, for several reasons.

First, the estimate of R over a chosen time window can only be obtained at the end of that window, since it requires the observation of incident cases over the whole window.

Moreover, we saw in section Web Appendix 2 that the precision of the estimate is higher for a higher number of incident cases. To get a posterior CV of 0.3 for example (which is the aimed CV we used to get the results presented in the main text), the time window considered must comprise at least 11 incident cases (see Web Table 1). We therefore advise starting estimating R only after 12 (the initial case + 11) cases have been observed at total. Anyway, there is little chance that an epidemic in its very early stage, with <12 cases, would be detected, unless the symptoms are extremely severe.

Finally, the infectiousness profile should also provide guidance on when to start estimation. Indeed, estimating R before at least one generation of cases has been observed is difficult. For example, in the extreme case where generations are discrete, all cases in the second generation are infected at the same time *t* after the index case: it is therefore impossible to estimate R before that time *t*. Moreover, early in an outbreak, a substantial fraction of incident cases may be imported, which we do not account for in this study. Waiting until at least one average serial interval has passed should reduce the associated bias in the early estimates of R.

Overall, we suggest starting estimating R once those three criteria are fulfilled: at least after τ, at least after one mean serial interval, and when at least 12 cases have been observed since the beginning of the epidemic.

Web Appendix 4. Uncertainty in the infectiousness profile

Estimates of the reproduction number are highly dependent on the choice of the infectiousness profile w_s. This can be approximated by the distribution of the generation time (i.e. time from the infection of a primary case to infection of the cases he/she generates) . However, times of infection are rarely observed and the generation time distribution is therefore difficult to measure. On the other hand, the timing of symptoms onset is usually known and such data collected in closed settings where transmission can reliably be ascertained (e.g. households) can be used to estimate the distribution of the serial interval (time between symptoms onset of a case and symptoms onset of his/her secondary cases). Therefore, in practice, we apply our method on data consisting of daily counts of symptoms onset and where the infectivity profile w_s is approximated by the distribution of the serial interval. However, this distribution can be poorly documented, especially early in the epidemic. Here, we provide a method to explicitly take into account the uncertainty in the serial interval distribution. To do so, we assume that the serial interval is Gamma distributed, and we allow its mean μ_{SI} and standard deviation (sd) σ_{SI} to vary according to truncated normal distributions. We sample $n_{SI} = 1000$ pairs of mean and

sd: $(\mu_{SI},\sigma_{SI})^{(1)},\ldots,(\mu_{SI},\sigma_{SI})^{(n_{SI})}$, by first sampling $\mu_{SI}{}^{(k)}$, and then sampling $\sigma_{SI}{}^{(k)}$ with the constraint that $\sigma_{SI}{}^{(k)} < \mu_{SI}{}^{(k)}$. This constraint ensures that the Gamma probability density function of the serial interval is null at $t = 0$. For each pair $(\mu_{SI},\sigma_{SI})^{(k)}$, we then sample, for each sliding window of length τ, $n = 1000$ realizations $R^{(k,1)},\ldots,R^{(k,n)}$ of R in its posterior distribution, conditional on $(\mu_{SI},\sigma_{SI})^{(k)}$, forming at total a sample of size $n \times n_{SI} = 1,000,000$ of the joint posterior distribution of R.

We illustrated this method on the 1918 pandemic flu in Baltimore (see main text for the results), for which we used an average mean serial interval of 2.6 days (sd 1.5, min 1, max 4.2), and an average standard deviation of 1.5 days (sd 0.5, min 0.5, max 2.5).

Web Appendix 5. Estimation of the case reproduction number

In order to compare the our approach with the WT approach, we also estimated the case reproduction number $R^c_{t,\tau}$ using the Wallinga and Teunis (WT) method for the five dataset analysed in the main text as well as for the simulation study (see below). We estimated $R^c_{t,\tau}$, the average number of secondary cases infected by individuals with symptoms onset occurring during the time period $[t-\tau+1;t]$. As in WT, we estimated the mean case reproduction number of individual *j* as: $R^c_{ind\ j} = \sum_i \frac{w_{t_i - t_j}}{\sum_{k,k\neq i} w_{t_i - t_k}}$, where t_i is the time of symptoms onset of individual *i*. The mean estimated case reproduction number over the time window $[t-\tau+1;t]$ was then estimated by averaging the individual case reproduction number over all those with symptoms onset in the considered time window: $R^c_{t,\tau} = \frac{\sum_{j;t-\tau+1\le t_j\le t} R^c_{ind\ j}}{\sum_{j;t-\tau+1\le t_j\le t} 1}$. The confidence intervals were obtained by reconstructing a panel of possible infection trees using multinomial allocation of infectors. All the results presented here were obtained with 100 reconstructed trees. Note that no estimates are available for weeks with no incident cases.

For the simulation study, we derived the theoretical case reproduction number from the instantaneous reproduction number using the following formula: $R^c_t = \sum_{s=0}^{+\infty} R_{t+s} w_s$.

Web Appendix 6. Simulation study

In order to assess the ability of our method to quantify transmissibility in several epidemic scenarios, we designed a simulation study based on two scenarios:

- constant instantaneous reproduction number $R = 2.5$,
- constant *R* before ($R = 2.5$) and after ($R = 0.7$) a certain date, illustrating the effect of a control measure such as school closure.

For each scenario, we simulated 100 epidemics, starting with 10 index cases. We used a SARS like serial interval distribution, with mean 8.4 days and standard deviation 3.8 days. We assumed a simple scenario with constant incubation period, so that the incidence of symptomatic cases is exactly the incidence of infections, but shifted in time. For each day $t \geq 2$, the number of incident cases I_t was drawn from a Poisson distribution with mean $R_t \sum_{s=1}^{t} I_{t-s} w_s$, where w_s is the discrete serial interval distribution. The epidemics were run for $T = 50$ days, with the intervention in scenario 2 occurring on day $T_e = 15$. For each simulated epidemic, we then reestimated, using our method, the instantaneous reproduction number $R_{t,\tau}$.

In order to compare the two approaches, we also estimated the case reproduction number $R^c_{t,\tau}$ using the Wallinga and Teunis (WT) method (see previous section) .

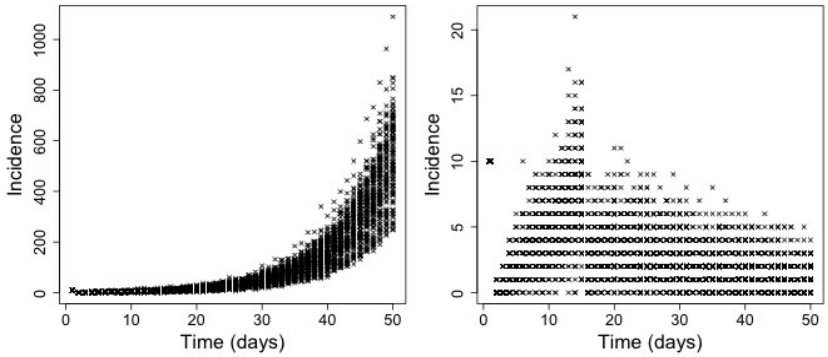
The simulated epidemics are shown in Web Figure 1 and the corresponding estimated reproduction numbers shown in Web Figures 2 and 3.

Our method allows reestimating the instantaneous reproduction number used for simulation. Unlike the WT method, it does not suffer of the right censoring, i.e. estimates of the reproduction number at the very end of the time series accurately reflect the transmissibility at that time point and do not artefactually decrease to zero due to lack of observation of secondary cases in the future (see Web Figure 2). However it's worth mentioning that methods inspired from the WT method have been developed to overcome this issue .

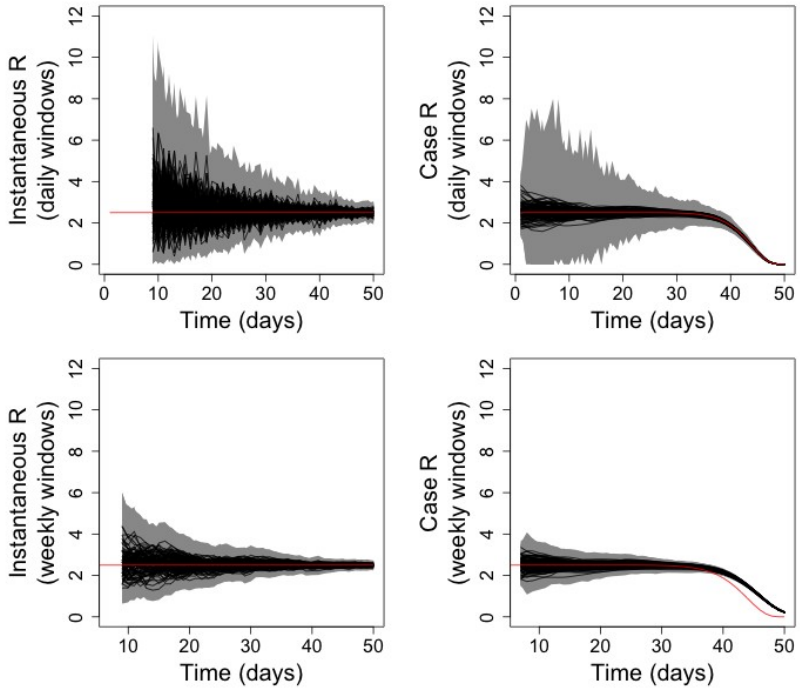
The case reproduction number $R^c_{t,1}$ on a day t reflects transmissibility over a time period starting on day t and lasting for one serial interval, whereas the instantaneous reproduction number $R_{t,1}$ on day t reflects transmissibility on a single day. Therefore $R^c_{t,1}$ is smoother than $R_{t,1}$. Similarly for any given window size τ, the instantaneous *R* over that window estimated with our method is more variable from one time window to the next than the corresponding case R^c (see Web Figure 2).

Interestingly, our method allows detecting changes in the instantaneous reproduction number, for instance a decrease in transmissibility following a control measure. These features are more difficult to detect when estimating R^c because it is not an instantaneous measure of transmissibility, and therefore its variations are much smoother over time (see Web Figure 3).

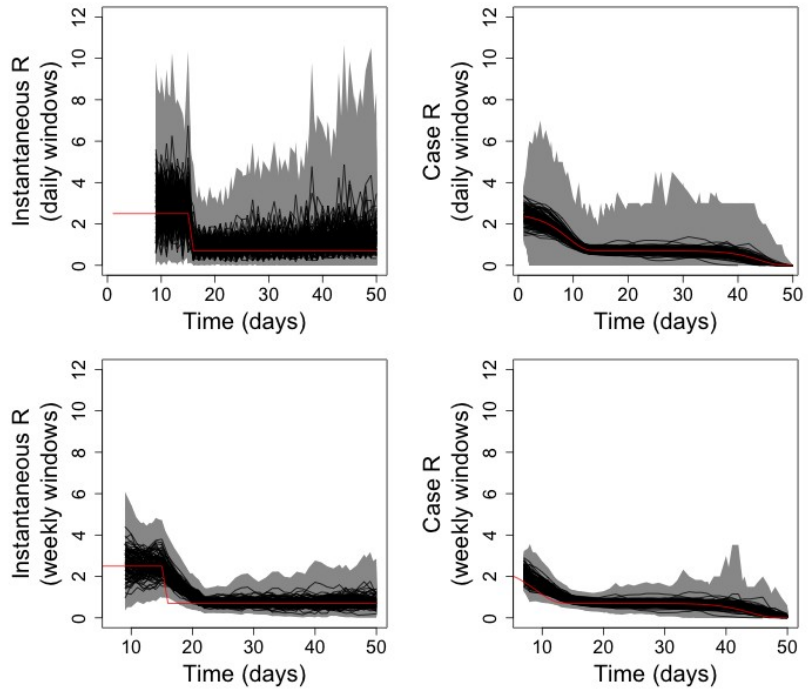
Using larger time windows allows getting smoother estimates of *R*, which leads to smoother but delayed curves and lowers the ability to detect changes in transmissibility. However, since our method is based on analytical estimates, analyses are fast and one can choose several time windows to analyse a dataset.



Web Figure 1 : Simulated epidemic curves in scenarios with constant instantaneous reproduction R=2.5 (left), constant R before (R=2.5) and after (R=0.7) a control measure on day 15 (right). 100 epidemics were simulated for each scenario, using a SARS-like serial interval.



Web Figure 2 : Instantaneous reproduction number (left panels) and case reproduction numbers (right panels) estimated, for 100 epidemics simulated under scenario 1 (constant transmissibility) by our method and the Wallinga and Teunis (WT) method respectively, on daily windows (top panels) and sliding weekly windows (bottom panels). The black lines show the mean estimates and the grey zones show the 95% credible (our method) or confidence (WT method) intervals. The red lines show the instantaneous reproduction number used for simulation (left panels) and the corresponding case reproduction numbers (right panels) calculated as in .



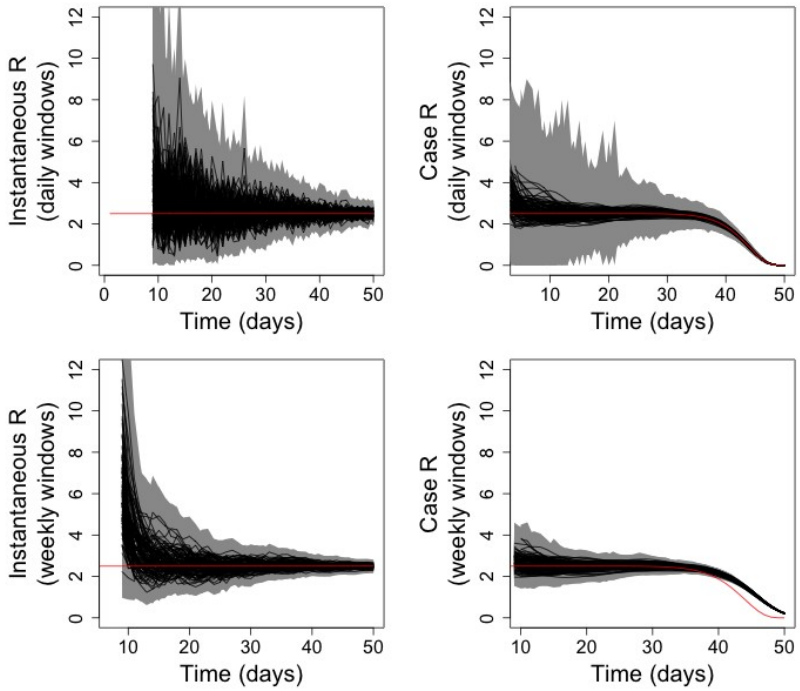
Web Figure 3 : Instantaneous reproduction number (left panels) and case reproduction numbers (right panels) estimated, for 100 epidemics simulated under scenario 2 (constant transmissibility before and after a control measure) by our method and the Wallinga and Teunis (WT) method respectively, on daily windows (top panels) and sliding weekly windows (bottom panels). The black lines show the mean estimates and the grey zones show the 95% credible (our method) or confidence (WT method) intervals. The red lines show the instantaneous reproduction number used for simulation (left panels) and the corresponding case reproduction numbers (right panels) calculated as in .

Web Appendix 7. Influence of incubation period distribution

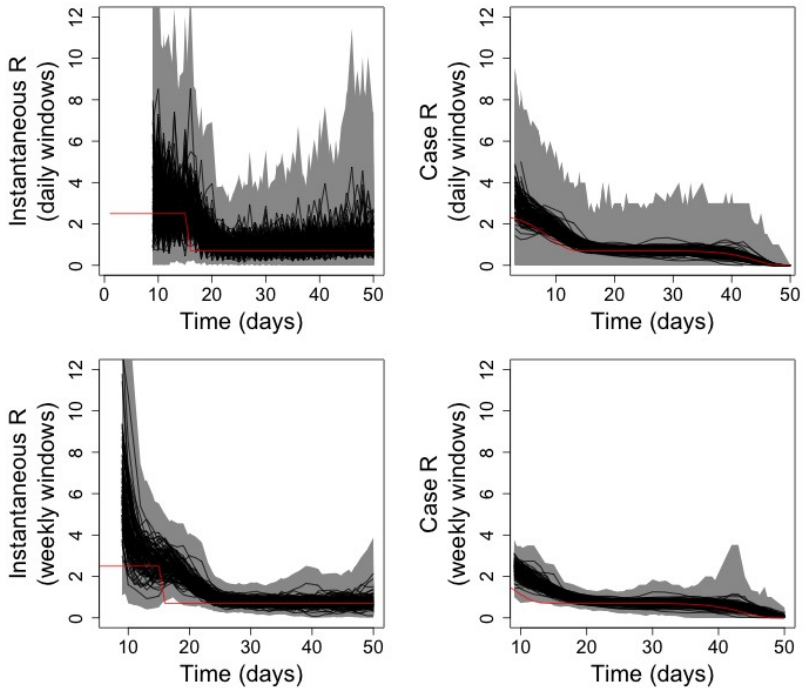
In the previous section, we assumed that the incubation period was constant, so that in scenario 2 all cases infected just after the control measure would have symptoms on the same day. If the incubation period is not constant, the effect will be diluted and the changes in the instantaneous reproduction number estimated from the times series of symptoms onset will be less abrupt.

In this section, we present a simulation study designed to assess whether changes in transmissibility would still be detected if the incubation period was variable amongst individuals.

We considered the simulated epidemics presented in the previous section, but where the dates are the dates of infection rather than symptoms onset. The incubation period for each case was then drawn according to a discretized Gamma distribution with mean 3.81 days and variance 8.34 $days^2$. The instantaneous and case reproduction numbers were then estimated from the times series of symptoms onset. Results are presented in Web Figures 4 and 5. Changes in transmissibility were still apparent in the estimates of the instantaneous reproduction number, but as expected, less clearly than in the scenario with constant incubation period. Generally speaking, the largest the variance in the incubation period, the lower the power to detect changes in transmissibility.



Web Figure 4 : Instantaneous reproduction number (left panels) and case reproduction numbers (right panels) estimated, for 100 epidemics simulated under scenario 1 (constant transmissibility) and assuming a non constant incubation period, by our method and the Wallinga and Teunis (WT) method respectively, on daily windows (top panels) and sliding weekly windows (bottom panels). The black lines show the mean estimates and the grey zones show the 95% credible (our method) or confidence (WT method) intervals. The red lines show the instantaneous reproduction number used for simulation (left panels) and the corresponding case reproduction numbers (right panels) calculated as in .



Web Figure 5 : Instantaneous reproduction number (left panels) and case reproduction numbers (right panels) estimated, for 100 epidemics simulated under scenario 2 (constant transmissibility before and after a control measure) and assuming a non constant incubation period, by our method and the Wallinga and Teunis (WT) method respectively, on daily windows (top

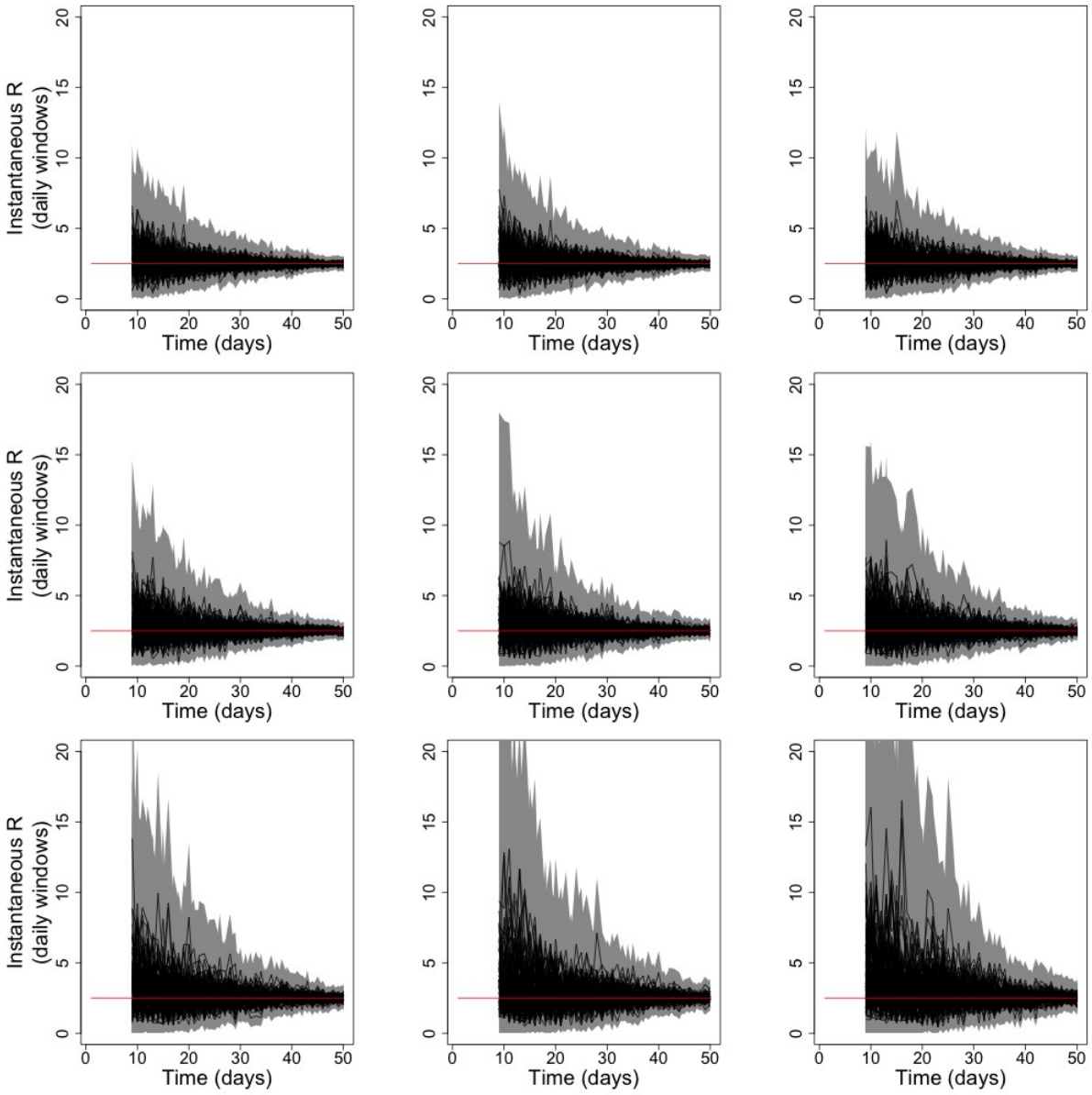
panels) and sliding weekly windows (bottom panels). The black lines show the mean estimates and the grey zones show the 95% credible (our method) or confidence (WT method) intervals. The red lines show the instantaneous reproduction number used for simulation (left panels) and the corresponding case reproduction numbers (right panels) calculated as in .

Web Appendix 8. Influence of underreporting

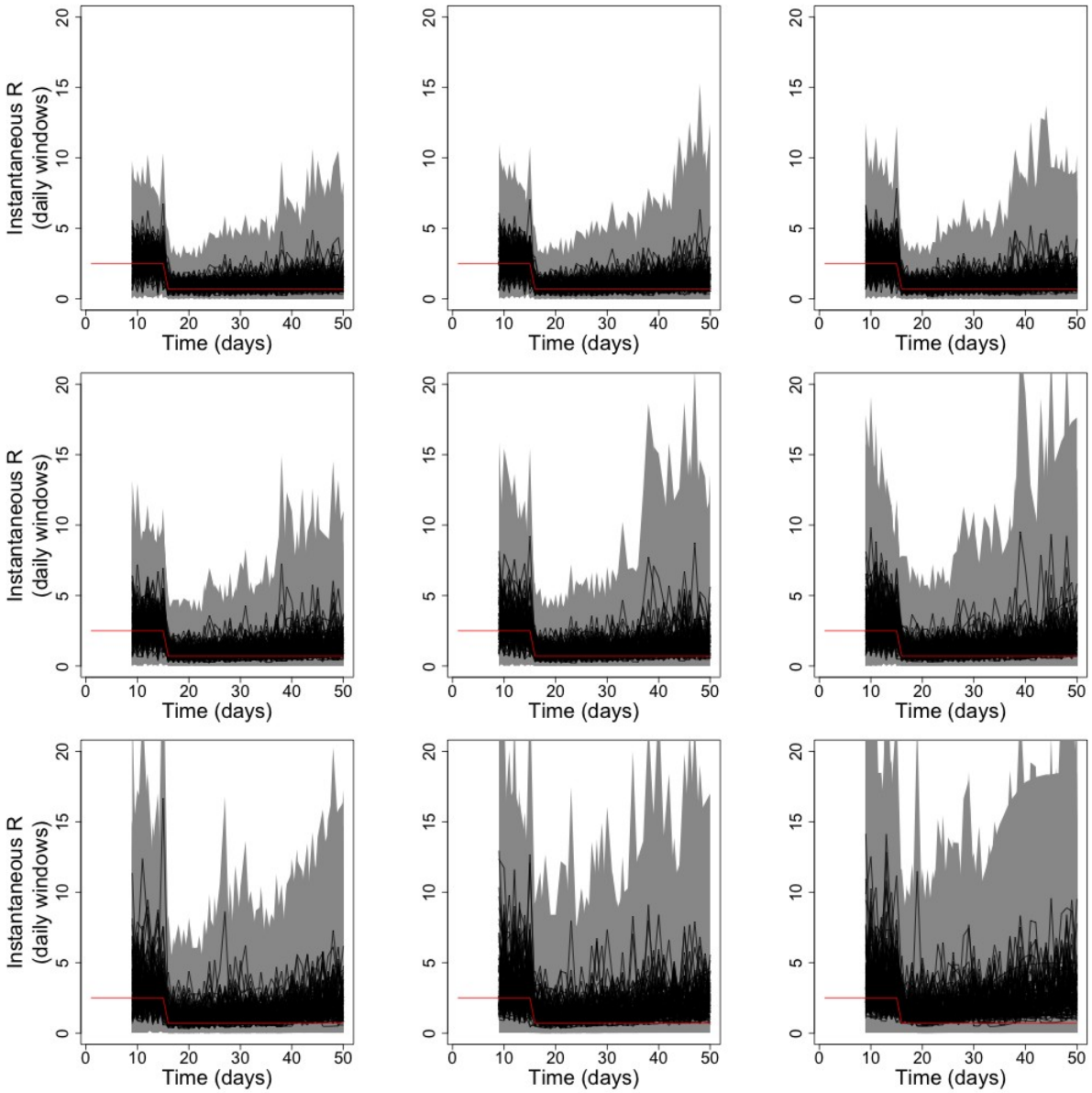
Simulations described in section 6 were further used to assess the influence of underreporting on the estimates of *R* obtained by using the time series of reported incidence. For each scenario, and for each of the 100 epidemics corresponding to that scenario, we simulated underreporting using a binomial distribution, with a constant reporting rate π varying between 20% and 100%. More precisely, for each day *t* of the epidemic, the number of reported cases O_t was drawn from a binomial distribution with parameters I_t (the true incidence on day t) and π. The instantaneous reproduction number was then reestimated crudely from O_t. The results are presented in Web Figures 6 and 7.

On average, the estimates of *R* from the reported cases only are similar to those obtained from all cases. However the credible intervals are wider for lower reporting rates (as expected since the number of cases in each time window is smaller, see section Web Appendix 2). Moreover, lower reporting rates lead to more variability in the mean estimates from one time window to the next, making it more difficult to detect changes in transmissibility.

Overall, underreporting does not appear to affect much the mean estimates of *R*, but affects the precision of those estimates.



Web Figure 6 : Instantaneous reproduction number R estimated on daily windows for 100 epidemics simulated under scenario 1 (constant transmissibility), with varying reporting rates (100%, 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20% respectively from top left to bottom right, line per line). The black lines show the mean estimates and the grey zones show the 95% credible intervals. The red lines show the R used for simulation.



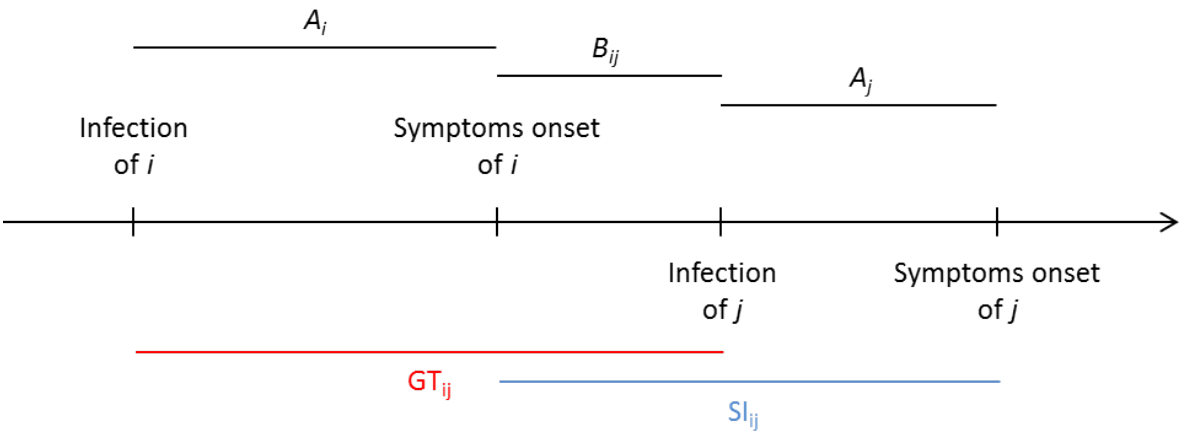
Web Figure 7: Instantaneous reproduction number R estimated on daily windows for 100 epidemics simulated under scenario 2 (constant transmissibility before and after a control measure), with varying reporting rates (100%, 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20% respectively from top left to bottom right, line per line). The black lines show the mean estimates and the grey zones show the 95% credible intervals. The red lines show the R used for simulation.

Web Appendix 9. Using symptoms onset and serial interval leads to exact estimates for diseases where infectiousness follows symptoms onset

Our approach to estimate the reproduction number is developed for the ideal situation where times of infection are known and the infectivity profile w_s may be approximated by the distribution of the generation time (time from the infection of a primary case to infection of the cases he/she generates) . However, surveillance data typically report times of symptoms onset rather than times of infection, and

as a consequence the generation time distribution is difficult to ascertain, unlike the serial interval distribution.

In this section, we consider diseases for which infectiousness only starts at or after the time of symptom onset. We propose a model similar to that used in Ferguson et al. . Each infected individual *i* experiences an incubation period A_i during which he/she is not symptomatic and not infectious. The incubation period is independently identically distributed in all individuals according to a distribution φ. The incubation period ends at the time of symptoms onset, and from that time, individual *i* has an infectiousness profile given by a distribution ψ, independent on the incubation period in individual *i*. If *i* infects an individual *j*, the sequence of infection and symptoms onset in *i* and *j* can be represented as in Web Figure .



Web Figure 8: Generation time and serial interval

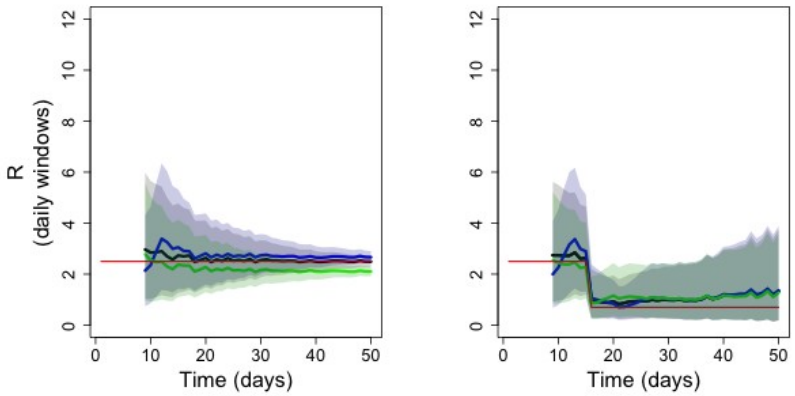
Note that in this model, B_{ij} is always positive as *i* is not infectious before symptoms onset. Moreover, the generation time and serial interval are given by $GT_{ij} = A_i + B_{ij}$ and $SI_{ij} = A_j + B_{ij}$ respectively. A_i, A_j and B_{ij} are independent, with respective distributions φ, φ and ψ. Hence the probability density functions of the generation time and the serial interval are given by $f_{GT} = \varphi * \psi$ and $f_{SI} = \varphi * \psi$ respectively, where $*$ is the convolution product, defined, for two probability density functions *f* and *g* by: $(f * g)(t) = \int_s f(t-s)g(s)ds$.

Therefore, in this case, the serial interval and the generation time have exactly the same distribution, and the estimates obtained using the serial interval distribution as the infectiousness profile will be exact. However, if the dates of symptoms onset are used instead of the dates of infection, the estimates are delayed, since R_t, estimated based on the new symptomatic individuals at time *t*, reflects in fact transmissibility at time $t - \delta$, where δ is the incubation period.

When the infectiousness profile ψ is not independent on the incubation period, the generation time and serial interval still have the same mean, but their variances can differ . To assess the extent to which

this could affect estimates of the instantaneous reproduction number obtained using the serial interval instead of the generation time, we reestimated *R* for the simulated epidemics described in section 6, but using a serial interval with standard deviation 2 times lower (respectively 2 times higher) than that used for the simulation (but same mean). Results are shown in Web Figure 9.

The instantaneous reproduction numbers reestimated with higher or lower standard deviation for the serial interval were able to capture changes in transmissibility. However, the mean estimates tended to be biased towards higher (respectively lower) values when a low (respectively high) standard deviation was used. In conclusion, using the serial interval distribution instead of the generation time distribution doesn't seem to affect the ability to detect changes in transmissibility, but overall, estimates of transmissibility might be slightly biased if the incubation period and the infectivity profile after symptoms are not independent.



Web Figure 9: Instantaneous reproduction number (*R*) reestimated from 100 simulated datasets under one of two epidemic scenarios (left: constant transmissibility, right: constant transmissibility before and after a control measure) over daily windows. The red lines show the instantaneous reproduction numbers *R* used for simulation. The black lines show the average (over the 100 simulated epidemics) of the mean *R* reestimated using the serial interval distribution used for simulation; the grey shaded areas are delimited by the average lower and upper bounds of the 95% credible intervals estimated using the serial interval distribution used for simulation. The blue (respectively green) lines show the average of the mean *R* reestimated using a serial interval with standard deviation two time lower (respectively two times higher) than that used for simulation ; the blue (respectively green) shaded areas are delimited by the average lower and upper bounds of the 95% confidence intervals estimated using a serial interval with standard deviation two time lower (respectively two times higher) than that used for simulation .

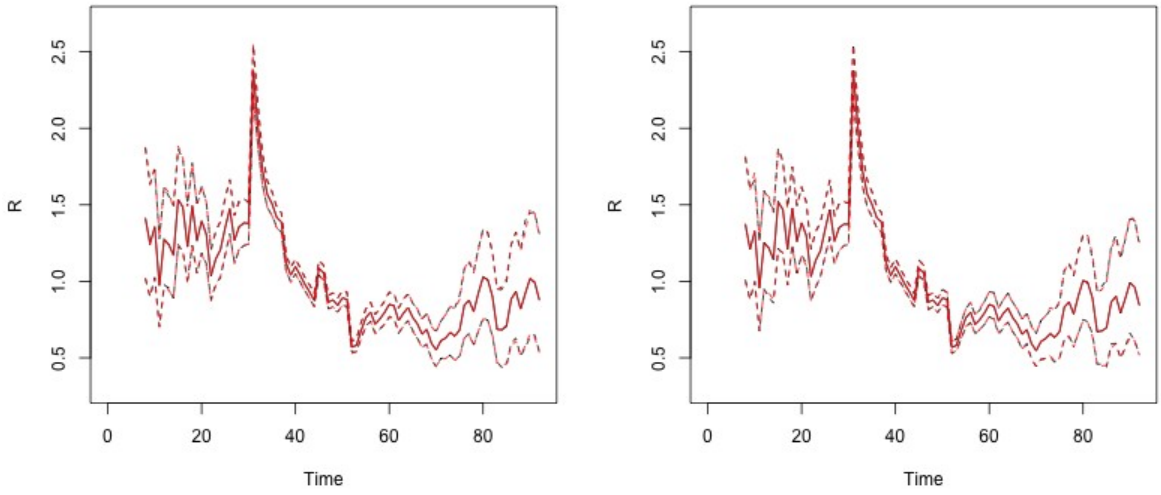
Web Appendix 10. Influence of the prior on the mean and variance of the serial interval distribution

We took the example of the outbreak of pandemic flu in Baltimore in 1918 to assess how the choice of the prior distribution influences the estimates of R .

First, we assessed the influence of the prior mean and variance on the estimates of R. We explored four scenarios with respective prior means 1, 5, 10 and 50, and standard deviations equal to the means (hence CV=1). A pairwise comparison of the four scenarios showed that the median estimates and upper

and lower bounds of credible intervals for R differed by less than 5%, suggesting that the results were little sensitive to the choice of the prior mean and variance.

We also explored the impact on the estimates of R of changing the form of the prior distribution. To do so, we implemented a Monte Carlo Markov Chain (MCMC) procedure, with a simple Metropolis-Hastings algorithm to update the states of the Markov chain, to estimate *R* assuming a Weibull distributed prior. We compared the resulting *R* estimates with the analytical estimates obtained assuming a Gamma prior with same mean and standard deviation. The analysis was performed twice, with a prior mean of 1 and 5 respectively (and standard deviation equal to the mean). The estimates of *R* were very little sensitive to both the form of the prior distribution (Gamma or Weibull) and the mean prior (1 or 5), as shown in Web Figure 10.



Web Figure 10: Daily estimates of the reproduction numbers *R* over sliding weekly windows for pandemic influenza in Baltimore, 1918, for a mean prior of 5 (left panel) and 1 (right panel); the black lines (hidden) correspond to Weibull prior and the red lines to Gamma prior; posterior medians are shown in plain line and 95% credible intervals in dotted lines.

Web Appendix 11. Discretization of serial interval distributions

Incidence data are typically discrete, so that the serial interval distribution needed to analyze them is discrete as well. However, most serial interval distributions, fitted to observations of transmission events in households for instance, are continuous. Here, we propose a formula to discretize the serial interval distribution.

We assume that the exact (i.e. continuous) time of infection of an incident case from day *t* is uniformly distributed on $\left[t;t+1\right[$. It can be shown that the delay *u* between the true times of infection of two cases that are incident on days *t* and *t+k* (*k≥0*) respectively is therefore distributed according to $f_U(u) = 1_{k-1<u<k+1}\left[1-\left|u-k\right|\right]$.

To discretize the serial interval distribution, we weight its probability density function with the probability function of each delay:

$$w_k = \int_{k-1}^{k+1} f_{SI}(u) f_U(u)\,du$$

$$= (1+k)F_{SI}(k+1) - 2kF_{SI}(k) + (k-1)F_{SI}(k-1) + \int_{k-1}^{k} u f_{SI}(u)\,du - \int_{k}^{k+1} u f_{SI}(u)\,du$$

which sum to 1.

- **Shifted Gamma distribution**

Assuming the serial interval SI is such that $\mathrm{SI}-1$ is Gamma distributed with probability density function $\mathrm{f}_{\mathrm{SI}-1}(\mathrm{t})=\frac{1_{\mathrm{t}\geq 0}}{\Gamma(\mathrm{a})\mathrm{b}^{\mathrm{a}}}\mathrm{t}^{\mathrm{a}-1}\mathrm{e}^{-\frac{\mathrm{t}}{\mathrm{b}}}$, we find:

$$w_k = k * F_{\Gamma,a,b}(k) + (k-2) * F_{\Gamma,a,b}(k-2) - 2*(k-1) * F_{\Gamma,a,b}(k-1) + ab\left(2F_{\Gamma,a+1,b}(k-1) - F_{\Gamma,a+1,b}(k-2) - F_{\Gamma,a+1,b}(k)\right)$$

Where $F_{\Gamma,a,b}$ is the cumulative density function of a Gamma distribution with parameters (a,b).

This is the parameterization that was used for all analyses in the paper unless otherwise specified, and it is the parameterization implemented in both the Excel® (Microsoft Excel®, Redmond, WA) tool and the *R* package.

- **Shifted Weibull distribution**

Assuming the serial interval SI is such that $\mathrm{SI}-1$ is Weibull distributed with probability density function $\mathrm{f}_{\mathrm{SI}-1}(\mathrm{t})=1_{\mathrm{t}\geq 0}\frac{\mathrm{a}}{\mathrm{b}}\left(\frac{\mathrm{t}}{\mathrm{b}}\right)^{\mathrm{a}-1}\mathrm{e}^{-\left(\frac{\mathrm{t}}{\mathrm{b}}\right)^{\mathrm{a}}}$, we find:

$$w_k = k * F_{W,a,b}(k) + (k-2) * F_{W,a,b}(k-2) - 2*(k-1) * F_{W,a,b}(k-1) + ke^{-\left(\frac{k}{b}\right)^a} + (k-2)e^{-\left(\frac{k-2}{b}\right)^a} - 2(k-1)e^{-\left(\frac{k-1}{b}\right)^a} - \frac{b}{a}\left[\gamma\left(\frac{1}{a},\left(\frac{k}{b}\right)^a\right) + \gamma\left(\frac{1}{a},\left(\frac{k-2}{b}\right)^a\right) - 2*\gamma\left(\frac{1}{a},\left(\frac{k-1}{b}\right)^a\right)\right]$$

where $F_{W,a,b}$ is the cumulative density function of a Weibull distribution with parameters (a,b), and $\gamma(s,x)=1_{x>0}\int_0^x t^{s-1}e^{-t}dt$ is the incomplete Gamma function.

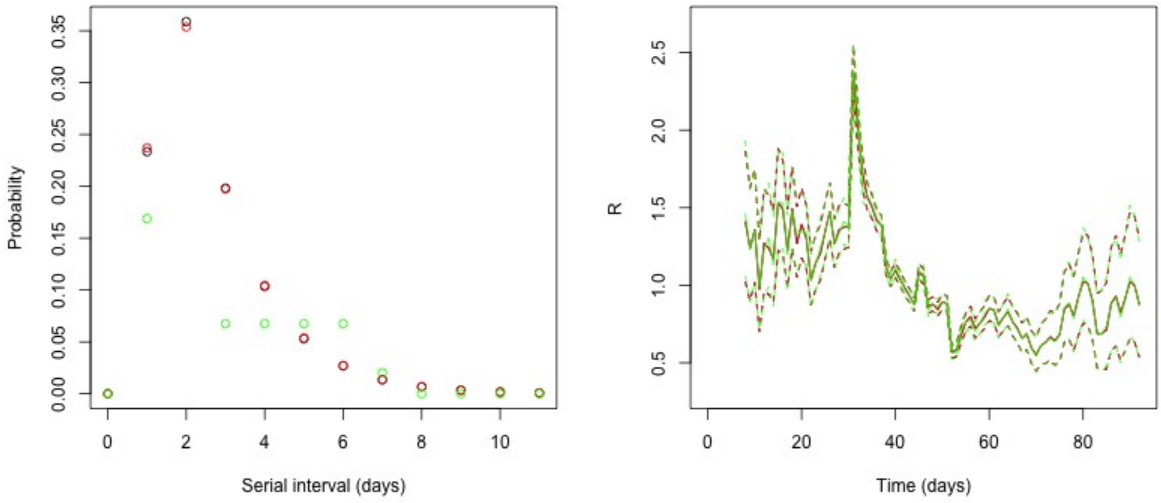
Web Appendix 12. Influence of the shape of the serial interval distribution

We took the example of the outbreak of pandemic flu in Baltimore in 1918 to assess how the shape of the serial interval distribution influences the estimates of R.

We explored three distributions for the serial interval:

- a discretized Gamma distribution with mean 2.6 days and standard deviation 1.5 day
- a discretized Weibull distribution with same mean and variance
- a manually constructed discrete distribution with same mean and variance but a different form.

Despite differences in the shape of the serial interval (especially between the manually constructed one and the two other ones), the estimates of *R* were very similar in all three scenarios, as shown in Web Figure 11.



Web Figure 11: Influence of the serial interval shape on the estimates of the reproduction number. Gamma (black), Weibull (red) and manually constructed (green) serial interval distributions with common mean and variance (left panel). Daily estimates of the reproduction numbers *R* over sliding weekly windows for pandemic influenza in Baltimore, 1918, using these three serial interval distributions (right panel); posterior medians are shown in plain line and 95% credible intervals in dotted lines.

Web Appendix 13. Derivation of the serial interval distribution for Measles and Smallpox

For all the outbreaks analysed in this article, we assumed a Gamma distribution for the serial interval, with mean and standard deviation taken from the literature. However, for measles and smallpox, the mean and standard deviation of the serial interval were not available directly from previous studies. They were derived indirectly from articles which reported related quantities such as the latency and the infectious period. In this section, we explain how to derive the serial interval distribution in a model with latency L and infectious period I. First we consider a model with constant infectiousness over the infectious period, and then we consider a model in which the infectious period is split in two periods with two different levels of infectiousness.

13.1. Measles: classical model with constant infectiousness over infectious period

Svensson shows that in such a model, the probability density function (pdf) of the serial interval (SI) is $f_{GI}(x) = \left[f_L * \left(\frac{1 - F_I}{\mathbf{E}[I]} \right) \right](x)$, where f_L is the pdf of the latency period L, F_I is the cumulative density function (CDF) of the infectious period I, $\mathbf{E}[I]$ is the average infectious period, and $*$ stands for the convolution operator.

The mean serial interval is therefore $\mathbf{E}[GT] = \mathbf{E}[L] + \frac{1}{\mathbf{E}[I]} \int_{x=0}^{+\infty} \left(1 - F_I(x)\right) x dx = \mathbf{E}[L] + \frac{\mathbf{E}[I^2]}{2\mathbf{E}[I]}$, and its variance is $\mathbf{Var}[GT] = \mathbf{Var}[L] + \frac{1}{\mathbf{E}[I]} \int_{x=0}^{+\infty} \left(1 - F_I(x)\right) x^2 dx - \left(\frac{\mathbf{E}[I^2]}{2\mathbf{E}[I]}\right)^2$. In the case where I is Gamma distributed with parameters (a_I, b_I) ($\mathbf{E}[I] = a_I b_I$ and $\mathbf{Var}[I] = a_I b_I^2$), the variance of the serial interval is therefore equal to:

$$\mathbf{Var}[GT] = b_I^{\,2}\left(\frac{1}{12}a_I^{\,2} + \frac{1}{2}a_I + \frac{5}{12}\right) + \mathbf{Var}[L]$$

In order to estimate the mean and variance of the serial interval for measles, we used these formulae together with the estimates of the average latency period (10.3 days) and of the standard deviation of the latency period (2.6 days) and the assumptions on the infectious period (leading to an average infectious period of 8.9 days and a standard deviation of 1.7 days) found in Groendyke et al. .

13.2. Smallpox: model with infectious period split in two parts with constant infectiousness over each part

We now consider a model in which the infectious period is split in two successive periods F and R, with two different levels of infectiousness. We assume that infectiousness during the second part (R) is k times infectivity during the first part (F).

A similar reasoning than for the model with constant infectiousness shows that the pdf of the serial interval is: $f_{GT}(x) = \left[f_L * \left(\frac{\left(1 - F_F\right) + k f_F * \left(1 - F_R\right)}{\mathbf{E}[F] + k\mathbf{E}[R]} \right) \right](x)$, where f_L is the pdf of the latency period L, F_F and F_R are the CDF of the first and second parts of the infectious period respectively, $\mathbf{E}[F]$ and $\mathbf{E}[R]$ are the corresponding means, and $*$ stands for the convolution operator.

After simplification, the mean serial interval is $\mathbf{E}[GT] = \mathbf{E}[L] + \frac{\mathbf{E}[F^2] + k\mathbf{E}[R^2] + 2k\mathbf{E}[R]\mathbf{E}[F]}{2\left(\mathbf{E}[F] + k\mathbf{E}[R]\right)}$

and its variance is:

$$\mathbf{Var}[GT] = \mathbf{Var}[L] - \left(\frac{\mathbf{E}[F^2] + k\mathbf{E}[R^2] + 2k\mathbf{E}[R]\mathbf{E}[F]}{2\left(\mathbf{E}[F] + k\mathbf{E}[R]\right)}\right)^2 + \frac{1}{3\left(\mathbf{E}[F] + k\mathbf{E}[R]\right)} \int_{t=0}^{+\infty} f_F(t) t^3 dt + \frac{k}{3\left(\mathbf{E}[F] + k\mathbf{E}[R]\right)} \left[\int_{u=0}^{+\infty} f_R(u) u^3 du + 3\mathbf{E}[R^2]\mathbf{E}[F] + 3\mathbf{E}[R]\mathbf{E}[F^2] \right]$$

In the case where F and R are Gamma distributed, the variance becomes

$$\mathbf{Var}[GT] = \mathbf{Var}[L] - \left(\frac{a_F b_F{}^2 (1+a_F) + k a_R b_R{}^2 (1+a_R) + 2k a_R b_R a_F b_F}{2(a_F b_F + k a_R b_R)} \right)^2 + \frac{b_F{}^3 (a_F+2)(a_F+1)a_F + k b_R{}^3 (a_R+2)(a_R+1)a_R + 3k a_R b_R a_F b_F \left[(1+a_R) b_R + (1+a_F) b_F \right]}{3(a_F b_F + k a_R b_R)}$$

In order to estimate the mean and variance of the serial interval for smallpox, we used these formulae together with the estimated durations of the latency (mean 11.6 days, variance 3.36 days^2), fever (mean 2.49 days, variance 0.89 days^2)and rash (mean 16.0 days, variance 18.3 days^2) periods, and the estimated relative infectiousness during the rash period compared to the fever period ($k = 6.4$) found in Riley and Ferguson .

Web Appendix 14. Guide for using the Excel® tool

In this section, we provide step-by-step guidance to using our Excel® tool to estimate instantaneous reproduction numbers from a time series of incidence and a serial interval distribution.

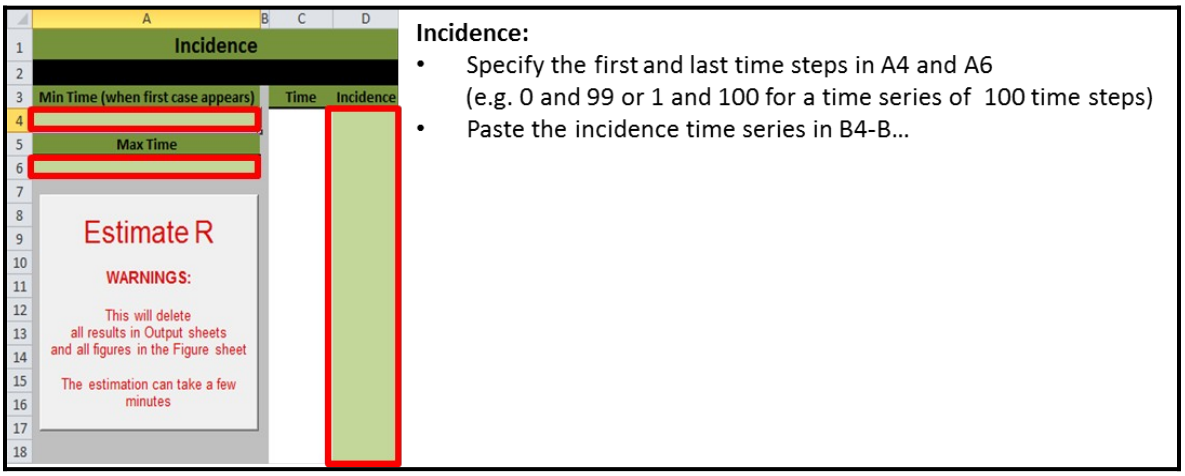
1. Open the Excel® file Estimation_R_instantaneous.xls

There are several sheets in it : Readme, Data, Output1 serial interval, Output2 R estimates and Figures. Readme will provide you with information on how to use the document, which is summarized in this document.

Data is the only sheet which needs to be modified; only light coloured cells have to be modified.

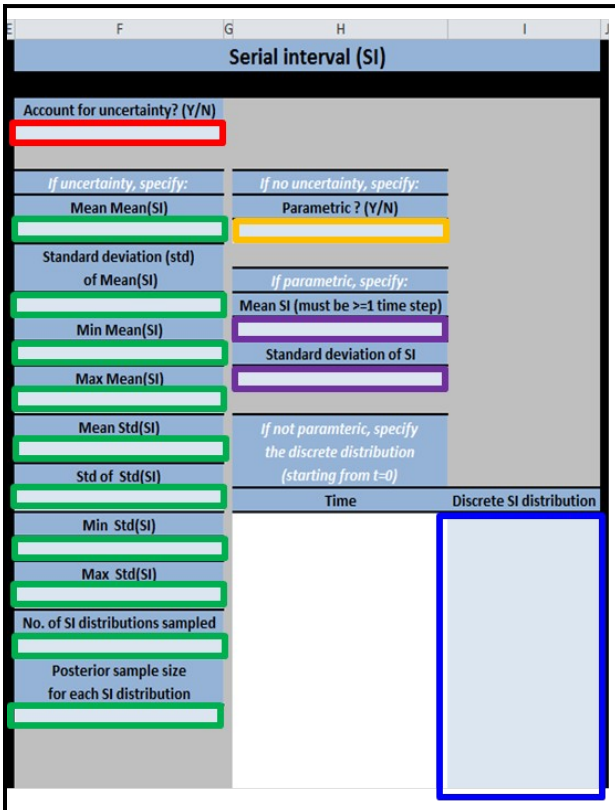
2. Fill in the Incidence section as shown in Snapshot 1.

Snapshot 1: Filling in the incidence section



3. Specify your assumptions about the serial interval distribution (see snapshot 2).

Snapshot 2: Specifying the serial interval distribution



Serial interval (SI):

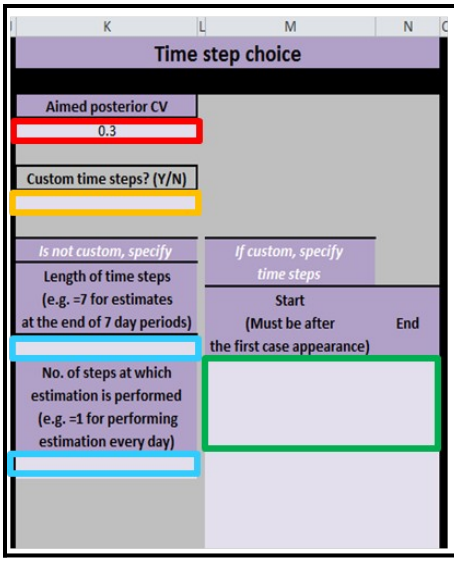
•Specify in F4 whether you want to account for uncertainty in the SI distribution (Y) or not (N)

•First option (N): not accounting for uncertainty
Specify the SI distribution either in a parametric or non parametric way (H8)
→If parametric, provide mean and sd for the SI (H12 and H14)
→If non parametric, provide (in I20-I…) the whole distribution of the SI (time step given by data, starting at t=0)

•Second option (Y): accounting for uncertainty
Provide Mean, Sd, Min and Max
- for the mean SI (F8, F11, F13, F15)
- for the sd of the SI (F17, F19, F21, F23)
Provide in F25 the number of SI distributions to explore (50 by default, the bigger the longer the estimation)
Provide in F28 the number of R values to be drawn for each SI distribution explored (50 by default, the bigger the longer the estimation)

4. Specify the time windows you want to use (see snapshot 3). Keep the posterior coefficient of variation to its default value of 0.3

Snapshot 3: Choosing the time windows for estimation of R



Time windows:

•Specify in K4 the aimed posterior coefficient of variation (default 0.3, the smaller the more precise the R estimates but the longer the time windows need to be to get this precision)

•Specify in K7 whether you will use custom time windows (Y) or not (N)

•First option (N): custom time windows

Give start times in M14-M… and end times in N14-N…

Time windows can overlap.

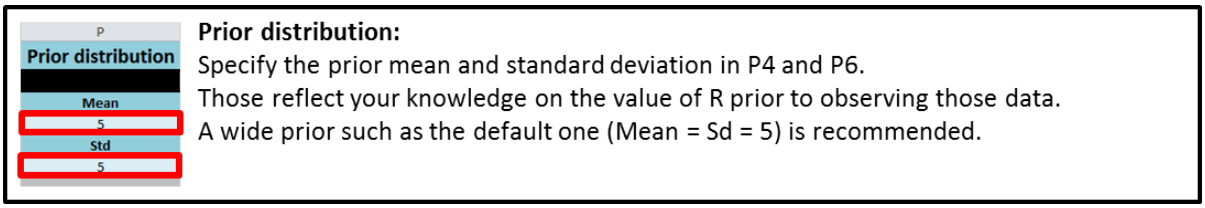
•Second option (Y): non custom time windows

Give the length of windows in K13 (eg 7 if daily data and weekly widows)

Give, in K18, the number of time windows at which estimation is performed (1 suggested to get estimates on sliding windows ending on each time step with data)

5. Specify the prior mean and standard deviation (see snapshot 4).

Snapshot 4: Specification of the prior distribution



6. Enable Macros

Macros need to be enabled in order to run the estimation of R. Refer to the documentation of Excel® specific to the version you are using to know how to enable the macros.

7. Run the estimation

Snapshot 5: Running the estimation

8. Reading the results

Results are presented as tables in sheets “Output1 serial interval” and “Output2 R estimates” and as figures in sheet “Figures”.

References

1. Fraser C. Estimating individual and household reproduction numbers in an emerging epidemic. PLoS ONE 2007;2(1):e758.
2. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol 2004;160(6):509-16.
3. Cauchemez S, Boelle PY, Thomas G, Valleron AJ. Estimating in real time the efficacy of measures to control emerging communicable diseases. Am J Epidemiol 2006;164(6):591-7.
4. Donnelly CA, Ghani AC, Leung GM, Hedley AJ, Fraser C, Riley S, et al. Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. Lancet 2003;361(9371):1761-6.
5. Ferguson NM, Cummings DA, Cauchemez S, Fraser C, Riley S, Meeyai A, et al. Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature 2005;437(7056):209-14.
6. Svensson A. A note on generation times in epidemic models. Math Biosci 2007;208(1):300-11.
7. Groendyke C, Welch D, Hunter DR. Bayesian Inference for Contact Networks Given Epidemic Data. Scandinavian Journal of Statistics 2011;38(3):600-616.
8. Riley S, Ferguson NM. Smallpox transmission and control: spatial dynamics in Great Britain. Proc Natl Acad Sci U S A 2006;103(33):12637-42.