PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Estimating the epidemic reproduction number from temporally aggregated incidence data: A statistical modelling approach and software tool

Rebecca K. Nash[1]*, **Samir Bhatt**[1,2], **Anne Cori**[1☯], **Pierre Nouvellet**[1,3☯]

1 MRC Centre for Global Infectious Disease Analysis, Jameel Institute, School of Public Health, Imperial College London, London, United Kingdom, **2** Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark, **3** School of Life Sciences, University of Sussex, Brighton, United Kingdom

☯ These authors contributed equally to this work.
* r.nash@imperial.ac.uk

Citation: Nash RK, Bhatt S, Cori A, Nouvellet P (2023) Estimating the epidemic reproduction number from temporally aggregated incidence data: A statistical modelling approach and software tool. PLoS Comput Biol 19(8): e1011439. https://doi.org/10.1371/journal.pcbi.1011439

Editor: Eric HY Lau, The University of Hong Kong, CHINA

Received: December 14, 2022

Accepted: August 18, 2023

Published: August 28, 2023

Data Availability Statement: We provide a GitHub repository containing the data and code to reproduce the analyses (https://github.com/rebeccanash/EM_EpiEstim_Nash2023) and used Zenodo to assign a DOI to the repository (10.5281/zenodo.8091058). The code to apply the method developed is available in the open-source R package EpiEstim: https://github.com/mrc-ide/EpiEstim.

Funding: RKN acknowledges funding from the Medical Research Council (MRC) Doctoral Training

Abstract

The time-varying reproduction number (R_t) is an important measure of epidemic transmissibility that directly informs policy decisions and the optimisation of control measures. EpiEstim is a widely used opensource software tool that uses case incidence and the serial interval (SI, time between symptoms in a case and their infector) to estimate R_t in real-time. The incidence and the SI distribution must be provided at the same temporal resolution, which can limit the applicability of EpiEstim and other similar methods, e.g. for contexts where the time window of incidence reporting is longer than the mean SI. In the EpiEstim R package, we implement an expectation-maximisation algorithm to reconstruct daily incidence from temporally aggregated data, from which R_t can then be estimated. We assess the validity of our method using an extensive simulation study and apply it to COVID-19 and influenza data. For all datasets, the influence of intra-weekly variability in reported data was mitigated by using aggregated weekly data. R_t estimated on weekly sliding windows using incidence reconstructed from weekly data was strongly correlated with estimates from the original daily data. The simulation study revealed that R_t was well estimated in all scenarios and regardless of the temporal aggregation of the data. In the presence of weekend effects, R_t estimates from reconstructed data were more successful at recovering the true value of R_t than those obtained from reported daily data. These results show that this novel method allows R_t to be successfully recovered from aggregated data using a simple approach with very few data requirements. Additionally, by removing administrative noise when daily incidence data are reconstructed, the accuracy of R_t estimates can be improved.

Partnership (grant reference MR/N014103/1). AC acknowledges the Academy of Medical Sciences Springboard, funded by the Academy of Medical Sciences, Wellcome Trust, the Department for Business, Energy and Industrial Strategy, the British Heart Foundation, and Diabetes UK (reference SBF005\1044). SB acknowledges support from the Novo Nordisk Foundation via The Novo Nordisk Young Investigator Award (NNF20OC0059309), The Eric and Wendy Schmidt Fund For Strategic Innovation via the Schmidt Polymath Award (G-22-63345), and the Danish National Research Foundation via a chair position. AC and SB acknowledge funding from the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Modelling and Health Economics, a partnership between the UK Health Security Agency, Imperial College London and LSHTM (grant code NIHR200908), and acknowledge funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement, and is also part of the EDCTP2 programme supported by the European Union. Disclaimer: The views expressed are those of the author(s) and not necessarily those of the NIHR, UK Health Security Agency or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: AC has received payment from Pfizer for teaching on a course for mathematical modelling of infectious disease transmission and vaccination.

Author summary

EpiEstim is a tool used to estimate the time-varying reproduction number, R_t, using only daily incidence data and the serial interval (SI) distribution—the estimated time between symptom onset in a case and their infector. R_t indicates whether case numbers are rising ($R_t > 1$) or falling ($R_t < 1$), with implications for disease control programmes. Frequently, incidence data are not reported daily, and when the SI of a disease is shorter than the temporal aggregation of incidence data, tools such as EpiEstim cannot be applied. Here, a novel method allows R_t to be estimated directly from reconstructed daily incidence data, using only aggregated incidence and the SI distribution. We validate our approach using influenza and COVID-19 data, alongside simulated incidence, to explore numerous epidemic scenarios. Trends in R_t estimated from reported daily incidence data can be recovered from daily incidence reconstructed from aggregated data. When administrative noise (e.g., weekend effects) is added to simulated incidence, R_t estimates from reconstructed data are more accurate, mitigating the impact of prominent 'noise' in daily reported incidence. Now implemented in EpiEstim, the method typically takes just a few seconds to run, making the tool applicable to a wider range of epidemic contexts.

Introduction

As infectious disease outbreaks become more common, it is increasingly important to rapidly characterise the threat of emerging and re-emerging pathogens [1]. Transmissibility, i.e. a pathogen's ability to spread through a population, can be quantified using the time-varying reproduction number, R_t, defined as the average number of infections that are caused by a primary case at time t of an outbreak. R_t signals whether an outbreak is growing ($R_t > 1$) or declining ($R_t < 1$), and whether current interventions are sufficient to control the spread of the disease.

One of the most popular tools for real-time R_t estimation, the R package EpiEstim, relies on observing the incidence data and supplying an estimated serial interval (SI) distribution—the time between symptom onset in a case and their infector. EpiEstim requires that the SI distribution and incidence data are supplied using the same time units. This can be problematic when daily incidence data is not reported, which is common for many diseases, such as influenza, Zika virus disease, and most notifiable diseases in countries such as the UK and the US [2–5]. Additionally, several studies intentionally aggregate data to reduce the impact of daily reporting variability; administrative noise, such as "weekend effects", are characterised by a drop in reported cases over weekends, due to reduced care seeking and longer delays in reporting, followed by a peak on Mondays [6,7]. A commonly used workaround is to aggregate the SI distribution to match the frequency of incidence reporting [8,9], however this is not possible if the SI is shorter than the aggregation of data. For example, influenza-like illness is typically reported on a weekly basis, but influenza has an estimated mean SI of 2–4 days [10,11]. Similarly, reporting of COVID-19, which has an estimated SI of 3–7 days, has typically moved from daily to weekly [12,13]. Therefore, enabling estimation of R_t from temporally aggregated data is critical to ensure methods such as EpiEstim are widely applicable [14].

In this study, we combine an expectation-maximisation (EM) algorithm with the renewal equation approach implemented in EpiEstim to reconstruct daily incidence from aggregated data and estimate R_t. We assess the performance of the method using influenza and COVID-19 data, in addition to an extensive simulation study.

Methods

EpiEstim

EpiEstim uses the renewal equation (Eq 1), a form of branching process model [15]. In this formulation, the incidence of new symptomatic cases at time t (I_t) is approximated by a Poisson process, where I_{t-s} is the past incidence, and g_s is the probability mass function of the serial interval.

$$I_t \sim Pois\left(R_t \sum_{s=1}^{t} I_{t-s} g_s\right) \tag{1}$$

With EpiEstim, R_t can be assumed to remain constant within user defined time windows, which smooth out estimates.

Extending EpiEstim for coarsely aggregated data

We extended EpiEstim to estimate R_t from aggregated incidence data, where each aggregation window (w) is >1 day, whilst still conditioning on an assumed serial interval distribution (g_s). We use an EM algorithm to iteratively reconstruct daily incidence from aggregated data, and in turn estimate R_t. We present the method with weekly data in mind, but the method and software can be applied to any temporal aggregation (Fig 1 and S1 Appendix pp 23–24).

We define:

- $\mathbf{I} = \{I_t\}_{t=1,\ldots,T}$ the vector of unobserved daily incidence,
- $\mathbf{A} = \{A_w\}_{w=1,\ldots,W}$ the vector of observed aggregated incidence, so that for each aggregation window w $A_w = \sum_{t=t_{w-1}+1}^{t_w} I_t$
- $\mathbf{R}^* = \left\{R_w^*\right\}_{w=1,\ldots,W}$ the vector of reproduction numbers corresponding to each incidence aggregation window. The $*$ indicates that this is only used in the EM algorithm to reconstruct the daily incidence **I**, and is distinct from the final estimated **R**.

We use the following indexes:

- t for days ($t = 1, \ldots, T$),
- w for aggregation windows ($w = 1, \ldots, W$),

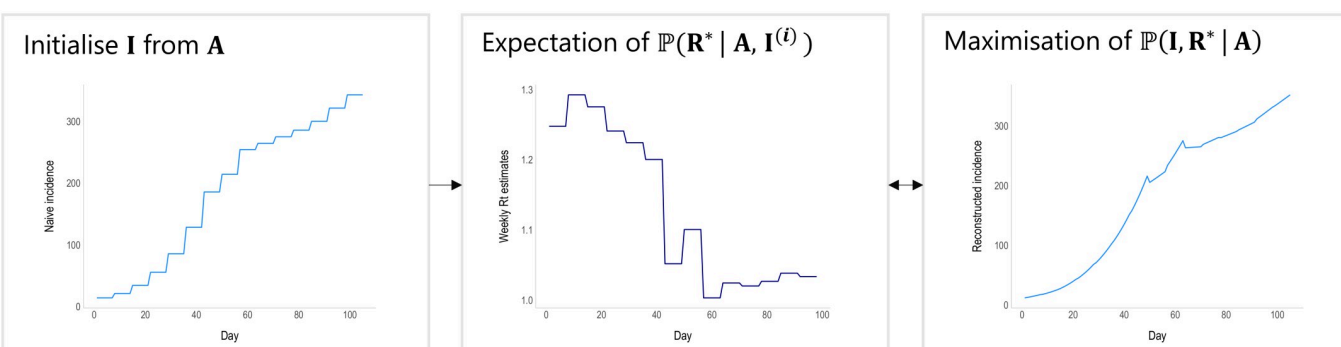

Fig 1. Schematic of the EM algorithm approach used to reconstruct daily incidence (**I**) from temporally aggregated incidence data (in this case weekly, **A**). The algorithm is initialised with a naive disaggregation of the weekly incidence (assuming constant daily incidence throughout the aggregation window, left panel). The resulting daily incidence is then used to estimate the reproduction number for each aggregation window, in this case for each week, $\mathbf{R}^*$ (expectation step, central panel). $\mathbf{R}^*$ is converted into a growth rate (see Eq 7), which is in turn used to reconstruct daily incidence data, whilst ensuring that if **I** were to be reaggregated it would still sum to the original weekly totals (maximisation step, right panel). The process cycles between the expectation and maximisation steps until convergence.

https://doi.org/10.1371/journal.pcbi.1011439.g001

- i for iterations of the EM algorithm ($i = 0, \ldots, 10$).

In the following, the bold notation signifies vectors. Our overall goal is to maximise the marginal likelihood function

$$\mathbb{P}(\mathbf{I}|\mathbf{A}) = \int \mathbb{P}(\mathbf{I},\ \mathbf{R}^*|\mathbf{A})d\mathbf{R}^* = \int \mathbb{P}(\mathbf{I}|\mathbf{R},\ \mathbf{A})\mathbb{P}(\mathbf{R}^*|A)d\mathbf{R}^*. \tag{2}$$

This marginal likelihood seeks to compute daily incidence while marginalising over the conditional probability distribution of the reproduction number. Loosely, the statistical goal is to produce a series of reproduction numbers that can reconstruct daily incidence while still being consistent with the observed aggregated incidence. We use an expectation maximisation scheme to approximate this marginal likelihood at low computational cost. The algorithm involves three steps: initialisation, expectation, and maximisation.

Initialisation

The algorithm is initialised (step $i = 0$) by disaggregating the aggregated incidence by piecewise constant functions (constant daily incidences across each aggregation window). That is, for aggregation window w covering days $t = t_{w-1} + 1, \ldots, t_w$:

$$I_t^{(i=0)} = \frac{1}{t_w - t_{w-1}} A_w. \tag{3}$$

Note that this allows non-integer incidence counts.

For iteration $i \geq 1$ of the algorithm, we iterate over two steps: expectation and maximisation.

Expectation

First, we compute the expectation

$$\mathbb{E}_{\mathbb{P}(\mathbf{R}^*|\mathbf{A},\ \hat{\mathbf{I}}^{(i)})}[\mathbb{P}(\mathbf{I},\ \mathbf{R}^*|\mathbf{A})]. \tag{4}$$

This expectation computes the average reproduction number over each data aggregation window, given the original observed aggregated incidence and the previous (i^{th}) iteration of the estimate of daily incidence. To compute this, we use the renewal approach from EpiEstim where the posterior distribution of reproduction numbers is found analytically as $\mathbb{P}(\mathbf{R}^*|\mathbf{I}) = \prod_{w=1}^{W} \mathbb{P}(R_w^*|\mathbf{I})$, with $\mathbb{P}\left(R_w^*|\mathbf{I}\right) \sim \text{Gamma}(\text{shape} = \alpha_w, \text{scale} = \beta_w)$ and where $\alpha_w = a + \sum_{s=t_{w-1}+1}^{t_w} I_s$ and $\beta_w = 1/\left(\frac{1}{b} + \sum_{s=t_{w-1}+1}^{t_w} \sum_{u=1}^{s} I_{s-u} g_u\right)$, where a and b are the shape and scale of the Gamma prior distribution for R_w.[15] The expected value for the reproduction number is therefore calculated as:

$$\hat{R}_w^{*(i)} = \frac{a + \sum_{s=t_{w-1}+1}^{t_w} \hat{I}_s^{(i-1)}}{\frac{1}{b} + \sum_{s=t_{w-1}+1}^{t_w} \sum_{u=1}^{s} \hat{I}_{s-u}^{(i-1)} g_u} \tag{5}$$

Maximisation

The maximisation step consists of recovering the most likely daily incidence from the expected $\mathbf{R}^*$ i.e. maximising $\mathbb{P}(\mathbf{I}, \mathbf{R}^* \,|\, \mathbf{A})$, or maximising $\mathbb{P}(\mathbf{I}, \mathbf{R}^*) \propto \prod_{t \geq 1} \mathbb{P}(I_t \,|\, I_0, \ldots, I_{t-1}, \mathbf{R}^*)$, subject to the constraint that daily incidence sums to the aggregated incidence i.e. $A_w = \sum_{t=t_{w-1}+1}^{t_w} I_t$.

In our renewal equation context, $I_t \,|\, I_0, \ldots, I_{t-1}, \mathbf{R}$ follows a Poisson distribution with mean $R_w \sum_s I_{t-s}\, g_s$ (where w is such that $t_{w-1} < t \leq t_w$), and therefore has mode $\hat{I}_t^{(i)} = \lfloor R_w \sum_{s=1}^{\infty} \hat{I}_{t-s}^{(i)} g_s \rfloor$

which we approximate as:

$$\hat{I}_t^{(i)} = \hat{R}_w^{*(i)} \sum\nolimits_{s=1}^{t} \hat{I}_{t-s}^{(i)} g_s. \tag{6}$$

Wallinga and Lipsitch [16] demonstrate, conditional on the generation time distribution, analytical correspondence between reproduction number R and growth rate r through the link function:

$$R = \frac{1}{\sum_{s-1}^{\infty} \exp(-rs) g_s} \tag{7}$$

We therefore assume local exponential growth so that Eq 6 is equivalent to:

$$\hat{I}_t^{(i)} = k_w e^{\hat{r}_w^{*(i)}(t-t_{w-1})}, \tag{8}$$

where $\hat{r}_w^{*(i)}$ is the exponential growth rate over aggregation window w, obtained from $\hat{R}_w^{*(i)}$ using the link function in Eq 7.

k_w is calculated to ensure the sum of daily incidence values adds up to the observed weekly totals:

$$k_w = \frac{A_w}{\sum_{t=t_{w-1}+1}^{t_w} \exp\left(\hat{r}_w^{*(i)}(t - t_{w-1})\right)} \tag{9}$$

We then use the estimate of **I** from Eq 8 in the maximisation step and iterate, thus completing the algorithm (Eq 5 for the expectation step and Eq 8 for the maximisation step). At this point, **I** can be used to estimate the full posterior distribution of **R** over any time window using EpiEstim (hereafter, this final **R** is referred to as R_t).

Given the rapid computational time and convergence (see S1 Appendix pp 10 and 22), the default number of iterations was set to 10 in the R package. However, a convergence check ensures that the final iteration of the reconstructed daily incidence does not differ from the previous iteration beyond a tolerance of 10^{-6}, and the number of iterations can be modified by the user.

Case studies

We chose datasets where incidence data was available daily, and then artificially aggregated them to weekly counts. R_t was estimated from daily incidence that was reconstructed from weekly aggregated data using our new approach, and compared to R_t estimates obtained from the reported daily incidence using the original EpiEstim R package. All R_t estimates were made using both daily and weekly sliding time windows, and we refer to those estimates as daily R_t estimates and weekly R_t estimates respectively.

We considered three characteristics: 1) mean R_t estimates, 2) uncertainty in the R_t estimates, and 3) the classification of R_t as increasing, uncertain or declining (S1 Appendix pp 8–9). To compare the performance of this approach to the original method, we assessed the correlations between each of the three characteristics when using the reported and reconstructed incidence. For the mean R_t estimates and uncertainty in R_t estimates, we assessed the linear relationships using the Pearson correlation coefficient (where values closer to +1 are indicative of a strong positive correlation).

The gamma distributed priors for $\mathbf{R}^*$ and R_t were set to a mean and standard deviation of 5 (shape = 1, scale = 5), which is the default prior parameterisation used in EpiEstim. The

rationale behind this choice is that it ensures that one will not conclude $R < 1$ unless the data strongly supports that. The user can set the prior themselves.

Influenza

We obtained a five-week subset of a dataset (11th December 2009—14th January 2010) on US active component military personnel (employed by the military as their full-time occupation) that made an outpatient visit to a permanent military treatment facility describing a respiratory-related illness. This daily incidence by date of presentation at a clinic was originally obtained by Riley et al. from the Armed Forces Health Surveillance Center and were digitally extracted for use here.[17] We used a mean SI of 3.6 days and SD of 1.6 days.[10]

COVID-19

Incidence of UK COVID-19 cases and deaths were taken from the UK government website [18]. For COVID-19 cases, we obtained ninety-seven weeks of data (21st February 2020 to 30th December 2021) for incidence by date of specimen, which is the date that a sample was taken from an individual which later tested positive. For COVID-19 deaths, we used ninety-six weeks of data (2nd March 2020 to 2nd January 2022) for incidence by date of death within twenty-eight days of a positive test. We assumed a mean SI of 6.3 days and SD of 4.2 days [12].

In the S1 Appendix, we also apply the EM algorithm to weekly incidence data for Zika virus disease to assess the performance of the method on a non-respiratory pathogen.

Simulation study

We considered scenarios where R_t either remained constant or varied over time, with a stepwise or gradual change. For each scenario, one hundred seventy-day epidemic trajectories were simulated using a Poisson branching process as implemented in the R package projections [19]. Daily datasets were aggregated weekly and used to estimate R_t using the proposed method; these values were compared to R_t estimates obtained from simulated daily data using the original EpiEstim R package. We explored the impact of weekend effects on R_t estimates, the ability to supply alternative temporal aggregations of data e.g., three-day, ten-day, or two-weekly aggregations, the ability to detect mid-aggregation variations in transmissibility, and finally, the number of iterations required to reach convergence when reconstructing daily incidence data. The full simulation study description and details can be found in S1 Appendix.

Results

Hereafter, we refer to reported and reconstructed incidence data, these are the reported daily incidence and the daily incidence that has been reconstructed from weekly aggregated data, respectively.

Influenza

The reconstructed incidence of influenza was much smoother than the reported incidence, which showed clear weekend effects and lower reported cases on two public holidays, both occurring on Fridays (Fig 2A and S1 Appendix p 8). Considering weekly sliding R_t first, there was a high correlation in both the mean R_t estimates derived from each dataset ($R^2 = 0.91$, Fig 2C and S1 Appendix p 2) and their associated uncertainty ($R^2 = 0.93$, Fig 2D). The overall agreement in the classification of R_t reached 81.8% (see methods and S1 Appendix p 9).

In contrast, mean daily R_t estimates differed markedly depending on whether the reported or reconstructed data were used, with an R^2 of 0.13 and much higher mean R_t and uncertainty

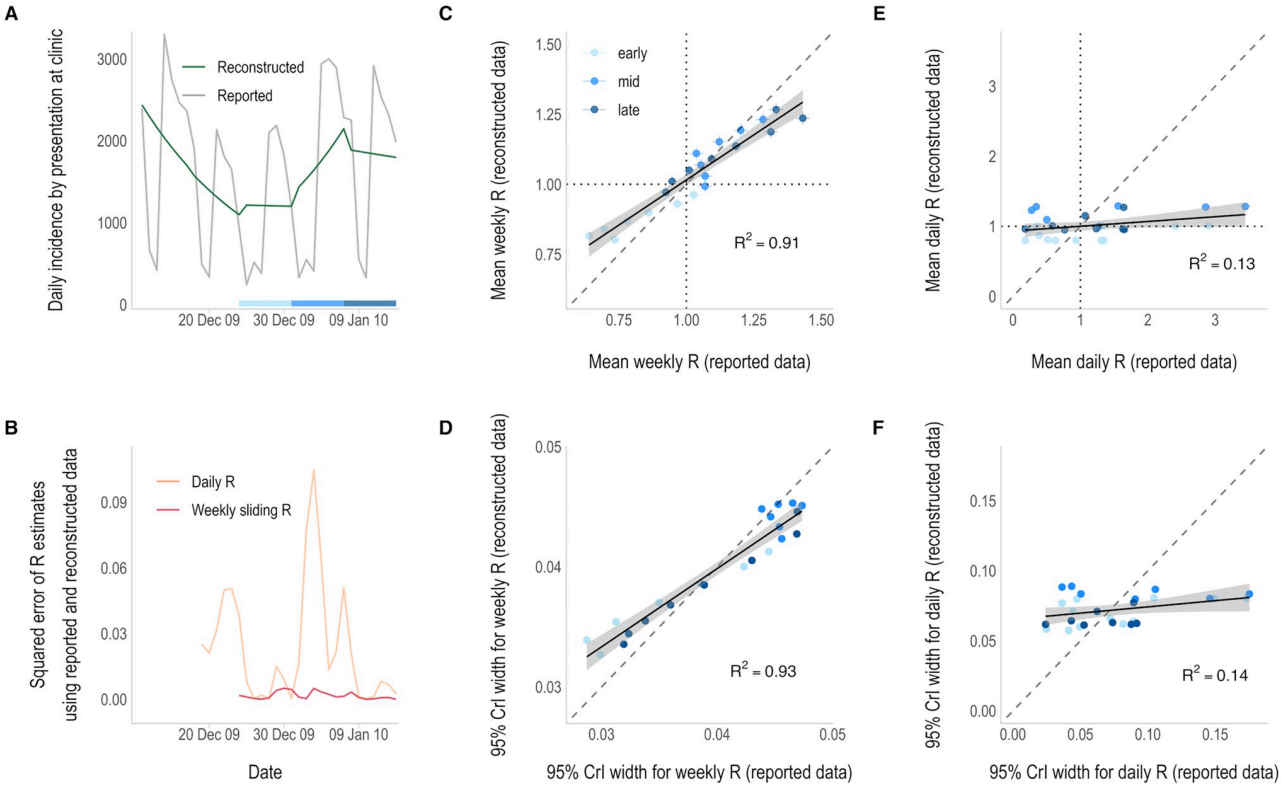

Fig 2. R_t estimates from daily incidence that was either reported or reconstructed from weekly aggregated influenza data. A) The reported (grey) and reconstructed (green) daily incidence of influenza by date of presentation at a military clinic. B) Squared error of the daily (orange) and weekly sliding (pink) R_t estimates that were made from reconstructed daily data compared to those obtained from the reported daily data. R_t estimation starts on the first day of the second aggregation window (day 8—18th December 2009) and is plotted on the last day of the time window used for estimation (i.e., starting on day 9 (19th December) for daily estimates and day 14 (24th December) for weekly estimates). Note: the x-axis is shared with the incidence plot above. C & E) Correlation between the weekly sliding (C) and daily (E) mean R_t estimates using reconstructed data (y-axis) and reported daily data (x-axis). Vertical and horizontal lines depict the 95% credible intervals (95% CrIs) and dotted lines show the threshold of $R_t = 1$. D & F) Correlation between the uncertainty in the weekly sliding (D) and daily (F) R_t estimates, defined as the width of the 95% credible intervals, using the reconstructed (y-axis) and reported (x-axis) daily data. The colour of the points in panels C-F correspond to the epidemic phase, i.e. the early (19th– 30th December for daily estimates, or 24th– 30th December for weekly sliding estimates), middle (31st December– 6th January) or late (7th– 14th January) phase of the data, shown by the strip in panel A. Solid lines show the linear model fit with 95% confidence intervals (grey shading). Dashed lines represent the x = y line.

https://doi.org/10.1371/journal.pcbi.1011439.g002

in estimates obtained from reported data (Fig 2E and 2F). Higher mean R_t estimates coincided with large peaks in the reported daily incidence (typically on Mondays), as daily R_t estimates were not smoothed and therefore more affected by intra-weekly variability (S1 Appendix p 2). The overall agreement in the classification of daily R_t estimates was much lower, with only 44.4% agreement (S1 Appendix p 9).

In this case study, the greatest differences in R_t estimates tended to correspond to time periods when the reported and reconstructed incidence data were most dissimilar (Fig 2B and S1 Appendix p 3). There was no apparent pattern in the estimates with regard to the outbreak phase, i.e. early, mid or late-phase, but this is likely due to this dataset being a snapshot of incidence taken from within an established epidemic (Fig 2).

COVID-19 cases

The reconstructed incidence of COVID-19 smoothed out intra-weekly variability, caused by factors such as weekend effects (Fig 3A and S1 Appendix pp 7–8). Weekly sliding R_t estimates

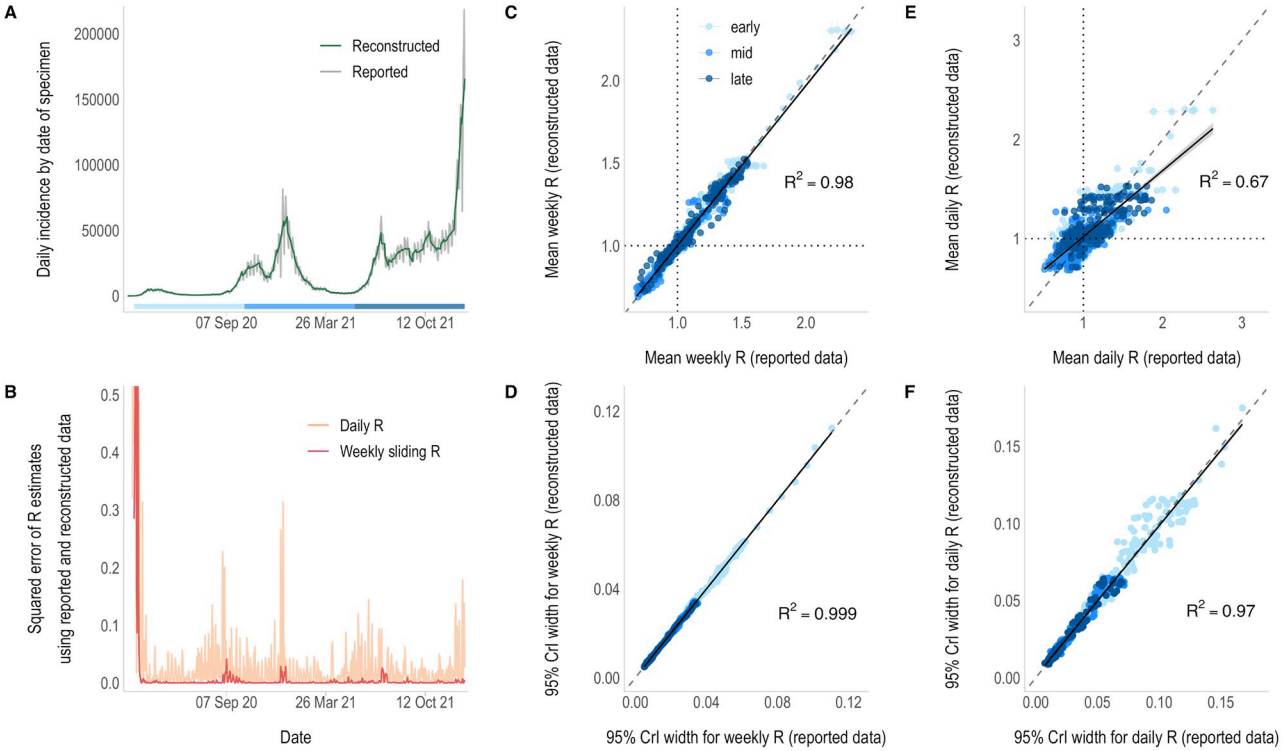

Fig 3. R_t estimates from daily incidence that was either reported or reconstructed from weekly aggregated COVID-19 case data. A) The reported (grey) and reconstructed (green) daily incidence of COVID-19 by date of specimen. B) Squared error of the daily (orange) and weekly sliding (pink) R_t estimates made from reconstructed data compared to those obtained from the reported daily data. R_t estimation starts on the first day of the second aggregation window (day 8—28th February 2020) and is plotted on the last day of the time window used for estimation (i.e., starting on day 9 (29th February) for daily estimates and day 14 (5th March) for weekly estimates). Note: the x-axis is shared with the incidence plot above and the y-axis has been limited to 0.5 for clarity. C & E) Correlation between the weekly sliding (C) and daily (E) mean R_t estimates using reconstructed (y-axis) and reported (x-axis) daily data, starting on day 30 due to low incidence. Vertical and horizontal lines depict the 95% credible intervals (95% CrIs) and dotted lines show the threshold of $R_t = 1$. D & F) Correlation between the uncertainty in the weekly sliding (D) and daily (F) R_t estimates, defined as the width of the 95% credible intervals, using the reconstructed (y-axis) and reported (x-axis) daily data. The colour of the points in panels C-F correspond to the epidemic phase, i.e. the early (21st March– 12th October 2020), middle (13th October 2020—22nd May 2021) or late (23rd May– 30th December 2021) phase of the data, shown by the strip in panel A. Solid lines show the linear model fit with 95% confidence intervals (grey shading). Dashed lines represent the x = y line.

https://doi.org/10.1371/journal.pcbi.1011439.g003

obtained from reconstructed and reported incidence were similar, both in their means (R^2 = 0.98) and their level of uncertainty (R^2 = 0.99, Fig 3C and 3D and S1 Appendix p 4). Mean daily R_t estimates were less well correlated (R^2 = 0.67), although the difference is less marked than in the influenza case study (Fig 3E), and the uncertainty in the estimates was similar across both approaches (R^2 = 0.97, Fig 3F). Most of the discrepant R_t estimates and higher levels of uncertainty coincide with the early phase of the outbreak when incidence was lower (Fig 3E and 3F). Outside of periods of low incidence, the largest differences in R_t estimates tended to correspond to time periods with greater disparities between the reported and reconstructed incidence data (Fig 3B and S1 Appendix p 5). The overall agreement in the classification of R_t estimates was higher than for influenza, with 74.4% and 94.9% agreement for daily and weekly sliding R_t estimates respectively (S1 Appendix p 9).

COVID-19 deaths

The reported incidence of COVID-19 deaths was much less influenced by day-to-day variation. The reconstructed daily incidence was more similar to the observed daily data than in the

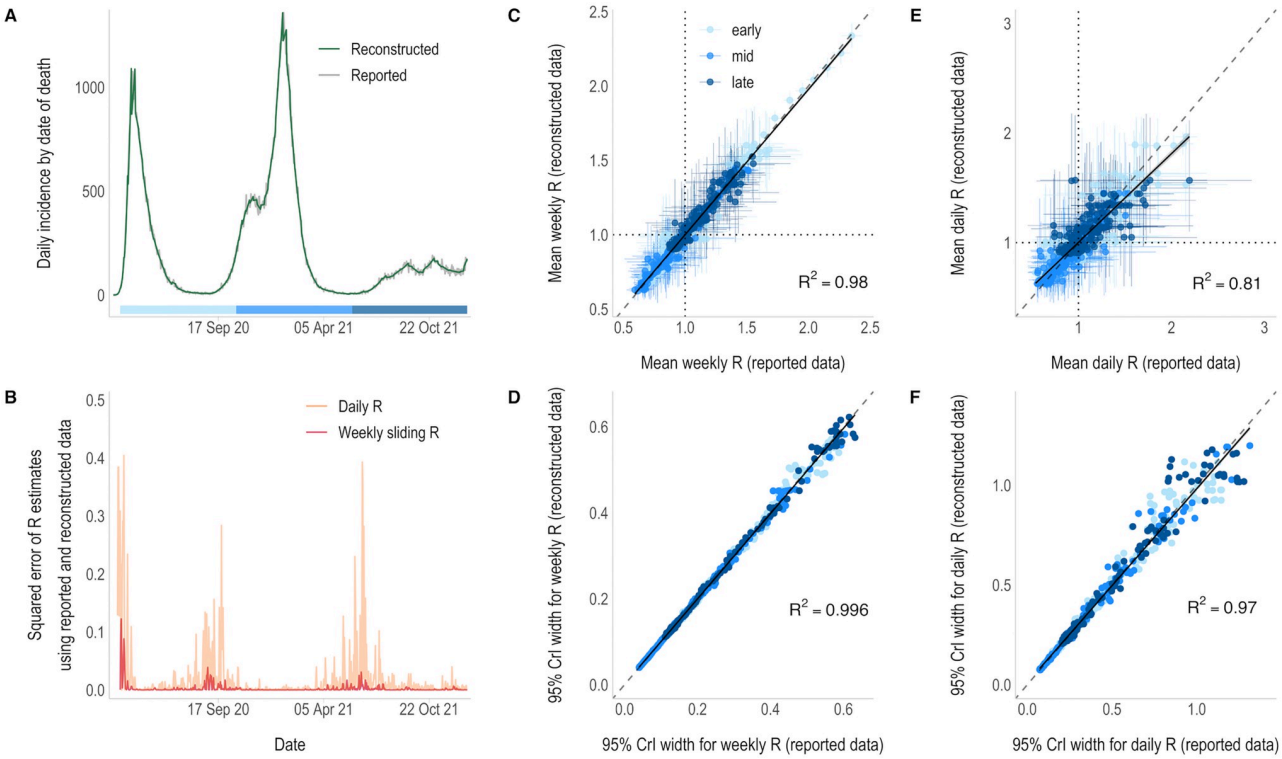

Fig 4. R_t estimates from daily incidence that was either reported or reconstructed from weekly aggregated COVID-19 death data. A) The reported (grey) and reconstructed (green) daily incidence of COVID-19 by date of death within 28 days of a positive test. B) Squared error of the daily (orange) and weekly sliding (pink) R_t estimates that were made from reconstructed data compared to those obtained from the reported daily data. R_t estimation starts on the first day of the second aggregation window (day 8—9th March 2020) and is plotted on the last day of the time window used for estimation (i.e., starting on day 9 (10th March) for daily estimates and day 14 (15th March) for weekly estimates). Note: the x-axis is shared with the incidence plot above and the y-axis has been limited to 0.5 for clarity. C & E) Correlation between the weekly sliding (C) and daily (E) mean R_t estimates using reconstructed (y-axis) and reported daily data (x-axis), starting on day 30 due to low incidence. Vertical and horizontal lines depict the 95% credible intervals (95% CrIs) and dotted lines show the threshold of $R_t = 1$. D & F) Correlation between the uncertainty in the weekly sliding (D) and daily (F) R_t estimates, defined as the width of the 95% credible intervals, using the reconstructed (y-axis) and reported daily (x-axis) data. The colour of the points in panels C-F correspond to the epidemic phase, i.e. the early (31st March–20th October 2020), middle (21st October 2020—28th May 2021) or late (29th May 2021—2nd January 2022) phase of the data, shown by the strip in panel A. Solid lines show the linear model fit with 95% confidence intervals (grey shading). Dashed lines represent the x = y line.

https://doi.org/10.1371/journal.pcbi.1011439.g004

previous case studies (Fig 4A). Both weekly and daily R_t estimates obtained from weekly data were highly consistent with those obtained from daily observations ($R^2 = 0.98$ and $R^2 = 0.80$ respectively, Fig 4C and 4E). The overall agreement in R_t classifications for daily estimates was the highest of all case studies at 85.8%, and 93.3% for weekly R_t estimates (S1 Appendix p 9). Discrepancies between the two mostly coincide with periods of particularly low incidence of deaths (Fig 4B and S1 Appendix p 7). The overall lower incidence of COVID-19 deaths compared to COVID-19 cases means there is greater uncertainty in R_t estimates in this case study (Fig 4D and 4F and S1 Appendix p 6). However, there was minimal difference in the uncertainty of estimates obtained from daily and weekly data (Fig 4D and 4F).

In all case-studies, incidence reconstructions converged within 10 iterations of the EM algorithm. The overall process of R_t estimation from weekly aggregated data took three seconds or less to run on MacOS (2 GHz Quad-Core Intel Core i5) 16GB RAM (S1 Appendix p 10); the influenza scenario, with over 57,000 cases, took two seconds to run, whilst the COVID-19 cases and deaths scenarios, with an overall incidence over 149,000 and 13 million cases respectively, took three seconds to run.

Simulation study

The method performed well across all scenarios, successfully estimating R_t from the aggregated simulated data, but unable to recover mid-aggregation window variations in transmissibility (S1 Appendix pp 10–24). Convergence of the EM algorithm was quick, with negligible differences in the reconstructed incidence beyond 5 iterations (S1 Appendix pp 22–23).

When introducing weekend effects into simulated data, R_t estimates from reconstructed incidence were more successful at recovering the true value of R_t than when using reported incidence (S1 Appendix pp 20–21). The method can also be successfully applied to other temporal aggregations of data, e.g. three-, ten- or fourteen-day windows (S1 Appendix pp 23–25).

Discussion

Estimates of the time-varying reproduction number (R_t) have frequently been used to inform and guide policymaking during outbreaks, and a commonly used approach to estimate R_t is EpiEstim, which relies on daily incidence data. However, maintaining daily incidence databases requires substantial time and investment in resources, which is not always feasible, particularly for less acute or routinely reported diseases. Therefore, in practice, many diseases are not reported on a daily basis, including influenza and other notifiable diseases in the UK and US [2–5]. As the COVID-19 pandemic persists, daily reporting is also becoming less common [20]. Coarsely aggregated data can be challenging to deal with in the context of R_t estimation methods, restricting their applications in certain contexts. In this study, we develop a statistical framework and tool that allows R_t estimation from aggregated incidence without introducing bias. Using influenza and COVID-19 data, alongside a simulation study, we demonstrate how a simple expectation-maximisation algorithm approach can rapidly reconstruct daily incidence data and accurately estimate R_t.

In all case studies, direct comparisons between weekly sliding R_t estimates show that very similar estimates can be made from the reported daily incidence and the reconstructed daily incidence from weekly aggregated data. However, daily R_t estimates are more influenced by noise, such as intra-weekly variability, leading to greater disparities in estimates between datasets. There are clear weekend effects exhibited in the influenza and COVID-19 case data (S1 Appendix p 8), leading to peaks and troughs in the reported incidence and the resulting daily R_t estimates (Figs 2 and 3, S1 Appendix pp 2 and 4). Using reconstructed incidence considerably smoothed the daily R_t estimates, removing the impact of weekend-effects. The overall agreement in the classification of R_t as increasing, uncertain, or declining between estimates made from each dataset rose substantially when some of the variability in the reported data was smoothed by estimating R_t using weekly sliding windows (S1 Appendix pp 8–9).

Despite both being affected by weekly periodicity in reporting, concordance of R_t estimates obtained from COVID-19 case data is considerably better than for influenza, perhaps due to the greater quantity of data, with a very strong positive correlation between daily and weekly R_t estimates (Fig 3). This is reflected in the high overall agreement in the classification of R_t estimates obtained from the reported and reconstructed datasets. It is important to note that outlying and much larger R_t estimates obtained from both datasets coincide with the early phase of the epidemic, when incidence was lower and the prior for R_t ($\mu = 5$, $\sigma = 5$) had more weight on estimates.

During the early stages of epidemics, despite there being far fewer deaths than cases, death data can sometimes be considered more reliable [21,22]. For example, case reporting is affected by surveillance system quality and the robustness of testing practices, which can vary considerably over the course of an epidemic, especially early on. COVID-19 incidence by date of death is much less influenced by administrative noise in the data (S1 Appendix p 8), and the

reconstructed incidence is most similar to the reported daily incidence of any case study. Therefore, the greatest differences in R_t estimates from death data coincide with periods of low incidence (S1 Appendix p 7) when uncertainty increases. Weekly sliding R_t estimates are equally as correlated as those from COVID-19 case data, but daily R_t estimates are the most strongly correlated of any dataset (Fig 4). Additionally, there is very high overall agreement in the classification of daily and weekly R_t (S1 Appendix p 9). This provides further support that differences between daily R_t estimates for influenza and COVID-19 cases is likely due to the reconstructed incidence smoothing out weekly periodicity in reporting.

To investigate further, weekend effects were artificially introduced to data in the simulation study (S1 Appendix p 20). We have shown that, when using reported incidence, R_t estimates are all strongly influenced by weekend effects (regardless of the smoothing time-window). Reconstructing daily incidence from weekly data completely removes the effect of noise from resulting R_t values, greatly improving the accuracy of estimates. This demonstrates that it may be beneficial to artificially aggregate daily data, as has been done in previous studies [6,7]. However, we did assume quite an extreme level of administrative noise, so in instances where the pattern is less prominent, it may have less of an impact on estimates. Furthermore, this smoothing effect could disguise genuine variations in transmissibility that occur mid-aggregation window, for instance, increased/decreased transmission over weekends (S1 Appendix section 3g). Disentangling important temporal trends in R_t from noise in the data can be difficult, and if aggregated data is used it will be at the cost of reduced temporal resolution in R_t estimates.

This can be seen when the method is applied to data aggregated over longer timescales, such as ten- to fourteen-days (S1 Appendix pp 23–25). This approach requires two layers of smoothing: 1) the incidence is smoothed over each aggregation window during the reconstruction process and 2) R_t estimates are smoothed by the sliding window chosen by the user. If a change in R_t occurs at the end of an aggregation window (i.e. on the last day), such as a sudden decrease in R_t due to a strict lockdown, that change is detected with a lag, corresponding to the length of the sliding window used for R_t estimation (S1 Appendix p 24). However, if the event occurs mid-aggregation window, then in addition to the usual lag caused by the sliding window, estimates will be affected by the smoothing of the incidence within the aggregation window during reconstruction (S1 Appendix p 25). The change in R_t will seem more gradual over the period that data are aggregated over and will appear to start earlier (corresponding to the first day of the aggregation window). It is important for users to be aware of this, particularly when using longer aggregations of data.

Another consideration is that the reconstructed incidence can have discontinuities in the borders between aggregation windows (S1 Appendix pp 11–12). This occurs because in reconstructing daily incidence we impose that, if it were to be re-aggregated, it would match the original data. Methods that simply fit smoothing splines to weekly data, inferring daily case counts from the daily difference in cumulative counts, are not affected by this [23,24]. To circumvent this problem, we recommend that sliding windows used to estimate R_t are at least equal to or longer than the length of aggregation windows to reduce the impact of discontinuities on estimates (S1 Appendix pp 23–25).

Alternative approaches include simple smoothing splines or LOESS to reconstruct daily incidence from aggregated data (see S1 Appendix section 4), and modelling frameworks implemented in the Epidemia and EpiNow2 R packages [6,21,25]. Daily infections are modelled as a latent process, back-calculated from observed data on cases or deaths, depending on an appropriate infection to observation distribution. In addition, Epidemia integrates further information, such as the infection ascertainment rate (for cases) or the infection fatality rate (for deaths) [21]. This facilitates a 'nowcasting' approach, allowing users to estimate R_t directly from the

unobserved infections, but they typically require more data (e.g. incidence of deaths and cases), more assumptions (e.g. delay distributions and ascertainment rates), and are much more computationally intensive, which can be a barrier to the adoption of such methods by users [14].

Here, R_t estimates are based on a single daily incidence reconstruction, meaning R_t can be estimated very rapidly from aggregated data, which is particularly desirable during real-time outbreak analysis [14]. A potential downside is that uncertainty in R_t estimates could be underestimated. However, the simulation study showed that the 95% credible interval of estimates encompassed the correct value of R_t the majority of the time, and we found no substantial indication that this approach detrimentally affected our characterisation of the uncertainty.

Given that this method is directly derived from EpiEstim, it relies on similar assumptions and caveats [15,26]. As time of infection is more difficult to observe than symptom onset, the SI is typically used as an approximation of the generation time in the renewal equation, which may introduce bias [27]. The SI, the level of undetected cases, and the reporting rate are assumed to remain constant, which is often not the case in practice. Factors such as changes in population immunity, and the introduction of interventions, can alter the SI throughout an epidemic [28]. Whilst changing case definitions, new testing practices, and increased healthcare-seeking behaviour, can all affect case ascertainment.[15] Parameters chosen by users can also influence estimation accuracy, for instance, the time window length for temporal smoothing and the prior for R_t [26]. Finally, EpiEstim's assumption of a Poisson likelihood may be a limitation in instances when data is substantially overdispersed [29,30].

To make the method simple to implement for current and future users of EpiEstim, this extension has been fully integrated with the 'estimate_R()' function in the original R package on GitHub [31]. Just one additional parameter is required—the number of days data are aggregated over (with some other optional parameters). The reconstructed daily incidence is also generated as an output, so it is possible to use it in other analysis pipelines involving alternative R estimation methods, which may perform better than EpiEstim in certain contexts, e.g. in retrospective analysis (S1 Appendix pp 29–30) [30], or in the presence of delays in reporting [25]. More details regarding the applications of this method can be found in the package vignette and associated examples [31].

Conclusion

We extended the widely used R_t estimation approach proposed by Cori et al., [15] and implemented in the R package EpiEstim, to incorporate a new feature which allows R_t to be easily estimated from any temporal aggregation of incidence data. We have demonstrated that the method performs well using both simulated and real-world data, recovering or even improving upon the estimates that would have been made from reported daily data. This extension is easy to use and computationally efficient, which will enable epidemiologists and other public health professionals to apply EpiEstim to a wider range of diseases and epidemic contexts.

Supporting information

S1 Appendix. Supplementary analyses.
(PDF)

Author Contributions

Conceptualization: Rebecca K. Nash, Anne Cori, Pierre Nouvellet.

Data curation: Rebecca K. Nash.

Formal analysis: Rebecca K. Nash.

Funding acquisition: Rebecca K. Nash.

Investigation: Rebecca K. Nash.

Methodology: Rebecca K. Nash, Samir Bhatt, Anne Cori, Pierre Nouvellet.

Project administration: Rebecca K. Nash.

Software: Rebecca K. Nash, Anne Cori.

Supervision: Anne Cori, Pierre Nouvellet.

Visualization: Rebecca K. Nash.

Writing – original draft: Rebecca K. Nash, Anne Cori, Pierre Nouvellet.

Writing – review & editing: Rebecca K. Nash, Samir Bhatt, Anne Cori, Pierre Nouvellet.

References

1. Baker RE, Mahmud AS, Miller IF, Rajeev M, Rasambainarivo F, Rice BL, et al. Infectious disease in an era of global change. Nat Rev Microbiol. 2022; 20: 193–205. https://doi.org/10.1038/s41579-021-00639-z PMID: 34646006
2. National flu and COVID-19 surveillance reports: 2021 to 2022 season. In: GOV.UK [Internet]. [cited 27 Jun 2022]. https://www.gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports-2021-to-2022-season
3. Pacheco O, Beltrán M, Nelson CA, Valencia D, Tolosa N, Farr SL, et al. Zika Virus Disease in Colombia—Preliminary Report. New England Journal of Medicine. 2020; 383: e44. https://doi.org/10.1056/NEJMoa1604037 PMID: 27305043
4. Notifiable diseases: weekly reports for 2022. In: GOV.UK [Internet]. [cited 27 Jun 2022]. https://www.gov.uk/government/publications/notifiable-diseases-weekly-reports-for-2022
5. Notifiable Infectious Disease Tables | CDC. 27 Sep 2021 [cited 2 Jul 2022]. https://www.cdc.gov/nndss/data-statistics/infectious-tables/index.html
6. Mishra S, Scott J, Zhu H, Ferguson NM, Bhatt S, Flaxman S, et al. A COVID-19 Model for Local Authorities of the United Kingdom. medRxiv; 2020. p. 2020.11.24.20236661. https://doi.org/10.1101/2020.11.24.20236661
7. Role of Data Aggregation in Biosurveillance Detection Strategies with Applications from ESSENCE. [cited 2 Jul 2022]. https://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a16.htm
8. Ferguson NM, Cucunubá ZM, Dorigatti I, Nedjati-Gilani GL, Donnelly CA, Basáñez M-G, et al. Countering the zika epidemic in latin america. Science. 2016; 353: 353–354.
9. Charniga K, Cucunubá ZM, Mercado M, Prieto F, Ospina M, Nouvellet P, et al. Spatial and temporal invasion dynamics of the 2014–2017 Zika and chikungunya epidemics in Colombia. PLOS Computational Biology. 2021; 17: e1009174. https://doi.org/10.1371/journal.pcbi.1009174 PMID: 34214074
10. Cowling BJ, Fang VJ, Riley S, Peiris JSM, Leung GM. Estimation of the serial interval of influenza. Epidemiology. 2009; 20: 344–347. https://doi.org/10.1097/EDE.0b013e31819d1092 PMID: 19279492
11. White LF, Wallinga J, Finelli L, Reed C, Riley S, Lipsitch M, et al. Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. Influenza and Other Respiratory Viruses. 2009; 3: 267–276. https://doi.org/10.1111/j.1750-2659.2009.00106.x PMID: 19903209
12. Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. The Lancet Infectious Diseases. 2020; 20: 911–919. https://doi.org/10.1016/S1473-3099(20)30287-5 PMID: 32353347
13. Rai B, Shukla A, Dwivedi LK. Estimates of serial interval for COVID-19: A systematic review and meta-analysis. Clin Epidemiol Glob Health. 2021; 9: 157–161. https://doi.org/10.1016/j.cegh.2020.08.007 PMID: 32869006
14. Nash RK, Nouvellet P, Cori A. Real-time estimation of the epidemic reproduction number: Scoping review of the applications and challenges. PLOS Digital Health. 2022; 1: e0000052. https://doi.org/10.1371/journal.pdig.0000052 PMID: 36812522

15. Cori A, Ferguson NM, Fraser C, Cauchemez S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. Am J Epidemiol. 2013; 178: 1505–1512. https://doi.org/10.1093/aje/kwt133 PMID: 24043437

16. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proceedings of the Royal Society B: Biological Sciences. 2007; 274: 599–604. https://doi.org/10.1098/rspb.2006.3754 PMID: 17476782

17. Riley P, Cost AA, Riley S. Intra-Weekly Variations of Influenza-Like Illness in Military Populations. Military Medicine. 2016; 181: 364–368. https://doi.org/10.7205/MILMED-D-15-00226 PMID: 27046183

18. Cases in the UK | Coronavirus in the UK. [cited 9 Jan 2022]. https://coronavirus.data.gov.uk/details/cases

19. Jombart T, Nouvellet P, Bhatia S, Kamvar ZN, Taylor T, Ghozzi S. projections: Project Future Case Incidence. 2021. https://CRAN.R-project.org/package=projections

20. CSSEGISandData. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. 2022. https://github.com/CSSEGISandData/COVID-19

21. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020; 584: 257–261. https://doi.org/10.1038/s41586-020-2405-7 PMID: 32512579

22. Nouvellet P, Bhatia S, Cori A, Ainslie KEC, Baguelin M, Bhatt S, et al. Reduction in mobility and COVID-19 transmission. Nat Commun. 2021; 12: 1090. https://doi.org/10.1038/s41467-021-21358-2 PMID: 33597546

23. Yamauchi T, Takeuchi S, Yamano Y, Kuroda Y, Nakadate T. Estimation of the effective reproduction number of influenza based on weekly reports in Miyazaki Prefecture. Scientific reports. 2019; 9: 1–9.

24. Nishiura H, Chowell G. Early transmission dynamics of Ebola virus disease (EVD), West Africa, March to August 2014. Eurosurveillance. 2014; 19: 20894. https://doi.org/10.2807/1560-7917.es2014.19.36.20894 PMID: 25232919

25. Abbott S, Hellewell J, Thompson RN, Sherratt K, Gibbs HP, Bosse NI, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Res. 2020; 5: 112. https://doi.org/10.12688/wellcomeopenres.16006.1

26. Gostic KM, McGough L, Baskerville EB, Abbott S, Joshi K, Tedijanto C, et al. Practical considerations for measuring the effective reproductive number, Rt. PLOS Computational Biology. 2020; 16: e1008409. https://doi.org/10.1371/journal.pcbi.1008409 PMID: 33301457

27. Britton T, Scalia Tomba G. Estimation in emerging epidemics: biases and remedies. Journal of The Royal Society Interface. 2019; 16: 20180670. https://doi.org/10.1098/rsif.2018.0670 PMID: 30958162

28. Ali ST, Wang L, Lau EHY, Xu X-K, Du Z, Wu Y, et al. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. Science. 2020; 369: 1106–1109. https://doi.org/10.1126/science.abc9004 PMID: 32694200

29. Brockhaus EK, Wolffram D, Stadler T, Osthege M, Mitra T, Littek JM, et al. Why are different estimates of the effective reproductive number so different? A case study on COVID-19 in Germany. medRxiv; 2023. p. 2023.04.27.23289109. https://doi.org/10.1101/2023.04.27.23289109

30. Gressani O, Wallinga J, Althaus CL, Hens N, Faes C. EpiLPS: A fast and flexible Bayesian tool for estimation of the time-varying reproduction number. PLOS Computational Biology. 2022; 18: e1010618. https://doi.org/10.1371/journal.pcbi.1010618 PMID: 36215319

31. mrc-ide/EpiEstim: A tool to estimate time varying instantaneous reproduction number during epidemics. [cited 9 Aug 2022]. https://github.com/mrc-ide/EpiEstim