Instantaneous Reproduction Number Estimation From Modelled Incidence

Robert Challen[1,2*], Leon Danon[1,2]

1 AI4CI, University of Bristol, Bristol, UK.
2 Department of Engineering Mathematics, University of Bristol, Bristol, UK.

* rob.challen@bristol.ac.uk

Abstract

The time varying reproduction number (R_t) is a critical quantity in monitoring an infectious disease outbreak. We propose a new method for estimating R_t from an infectivity profile expressed as a generation time distribution, and a time series of probabilistic estimates of disease incidence, modeled as log-normally distributed random variables. This is a common output of disease incidence models that are based on Poisson or negative binomial regression of case counts with a logarithmic link function. The method is deterministic, computationally inexpensive and it propagates uncertainty in the incidence estimates. It can be applied to estimates of disease incidence from a wide variety of incidence models, including those derived from weekly case counts. We validate the method when applied to the output of two simple statistical incidence models, and using simulated data with a defined R_t and infectivity profile. This combination produces comparable outputs to the de-facto standard 'EpiEstim'.

Author summary

In our experience estimating the reproduction number during the COVID-19 pandemic, we found that estimating the incidence rate was a useful first step to correct for artefacts in the raw count data, and to estimate the exponential growth rate. With incidence estimates available we wanted to use them to derive the time varying reproduction number to help monitor the state of the pandemic. We present a mathematical method and supporting software to estimate R_t from modelled incidence estimates, rather than raw count data, and which is readily applied to many incidence estimation models.

Introduction

Estimating the time varying reproduction number (R_t) is an important part of monitoring the progression of an epidemic, it informs short term projections of the epidemic size and hence guides decisions on policy interventions targetting behaviour [1]. Changes in R_t can reflect significant events in a pandemic such as the emergence of novel variants [2]. R_t estimation may be done using a range of techniques with varying degrees of complexity [3–15], but the majority are based on a time series of count data reflecting the incidence of infection in the population. Such count data may be new infections, hospitalisations or deaths, and are well known to exhibit specific biases due to incomplete ascertainment, reporting delay, and right truncation, along with more generic data quality issues such as missing values or anomalous values [16].

When correcting for such issues, one approach is to use count data to estimate incidence of a disease (I_t) using a model based around a time varying Poisson rate (λ_t), and fitted using maximum likelihood, using a logarithmic link function. In this common situation the estimate of the Poisson rate at any given time point (t) is a log-normally distributed quantity defined by parameters μ_t and σ_t, which include a representation of the uncertainty in the data. It is appealing to use such a modelled incidence estimate as the basis for an estimate of R_t, and include this uncertainty into R_t estimates. Incidence models can be derived in a number of ways, they are easily inspected for error and can be made tolerant of missing values and outliers.

A detailed review of the reproduction number is out of scope of this paper, and is well covered by Vegvari et al. [17], who highlight the difference between instantaneous and case reproduction numbers. The case reproduction number is the average number of secondary cases that arise from individuals infected today, and can only be estimated once these secondary cases have occurred. The instantaneous reproduction number estimates the number of primary infections in the past that have resulted in the secondary infections observed today, and hence is used for real time pandemic monitoring [1]. We concentrate solely on the instantaneous reproduction number in this paper. A key requirement for the estimation of the reproduction number is a profile of the delay between primary and secondary infections. This is described as the infectivity profile, a time dependent probability distribution, and is equivalent to the generation time distribution [1, 18]. The timing of sequential infections measured by the generation time is not directly observed, so the temporal distribution of the serial interval between the positive test results, or symptom onsets, of known infector-infectee pairs is often used as a proxy [6, 18]. The canonical framework for estimation of the instantaneous reproduction number direct from case data is the Cori method, as implemented in the R package 'EpiEstim' [6].

This paper presents a mathematical approach to estimating the instantaneous reproduction number from modelled incidence data, given an estimate of the infectivity profile, and which propagates incidence model uncertainty into estimates of the reproduction number. In this paper we refer to this method as 'R_t from incidence'. We validate the method against a simulation based on a branching process model with fixed infectivity profile and parameterised reproduction number and compare the output to reproduction number estimates using the Cori method implemented in the R package 'EpiEstim' [6].

Supporting implementations of all methods described here are provided in the associated R package "ggoutbreak" (https://ai4ci.github.io/ggoutbreak/).

Materials and methods

Mathematical analysis

To use a modelled estimate of incidence to predict R_t we need to propagate uncertainty in incidence into our R_t estimates. To calculate R_t we can use the backwards-looking renewal equations [1] which incorporate the infectivity profile of the disease (ω) at a number of days after infection (τ):

$$
\begin{aligned}
I_t &\sim Poisson(\lambda_t) \\
\lambda_t &\sim Lognormal(\mu_t, \sigma_t) \\
R_t &= \frac{I_t}{\sum_\tau \omega_\tau I_{t-\tau}}
\end{aligned}
\tag{1}
$$

If k is the length of the infectivity profile ($|\omega|$), in expectation, this gives:

$$R_t \approx \frac{\lambda_t}{\sum_{\tau=1}^{k} \omega_\tau \lambda_{t-\tau}}$$
$$\sim \frac{\text{Lognormal}(\mu_t, \sigma_t)}{\sum_{\tau=1}^{k} \text{Lognormal}\left(\mu_{t-\tau} + \log \omega_\tau, \sigma_{t-\tau}\right)} \tag{2}$$

It has been shown that the sum of correlated log-normal distributed random variables can be approximated by another log-normal [19] with parameters μ_Z and σ_Z, where the correlation between them is $\rho_{ij} = \text{Corr}(\log X_i, \log X_j)$.

$$S_+ = \mathbb{E}\left[\sum_i X_i\right] = \sum_i \mathbb{E}[X_i]$$
$$= \sum_i e^{\mu_i + \frac{1}{2}\sigma_i^2}$$
$$\sigma_Z^2 = \frac{1}{S_+^2} \sum_{i,j} \rho_{ij}\sigma_i\sigma_j \mathbb{E}[X_i]\mathbb{E}[X_j] \tag{3}$$
$$= \frac{1}{S_+^2} \sum_{i,j} \rho_{ij}\sigma_i\sigma_j e^{\mu_i + \frac{1}{2}\sigma_i^2} e^{\mu_j + \frac{1}{2}\sigma_j^2}$$
$$\mu_Z = \log S_+ - \frac{1}{2}\sigma_Z^2$$

The sum term in the denominator of the renewal equation (2) consists of a set of correlated scaled log normal distributions with scale defined by the infectivity profile (ω). For our case for a given time point t we equate $X_i = \omega_\tau \lambda_{t-\tau}$, and substitute $\mu_i = \mu_{t-\tau} + log(\omega_\tau)$ and $\sigma_i = \sigma_{t-\tau}$ into (3) to account for the infectivity profile. We define k to be the support of the infectivity profile ($k = |\omega|$). m_s is the weighted contribution from incidence estimates on day $t - \tau$, and Σ_{ij} is the covariance between log-incidence estimates from days $t - i$ and $t - j$.

$$m_\tau = e^{\mu_{t-\tau} + \log \omega_\tau + \frac{1}{2}\sigma_{t-\tau}^2}$$
$$\Sigma_{ij} = \text{Cov}(\log \lambda_{t-i}, \log \lambda_{t-j}) = \rho_{(t-i)(t-j)}\sigma_{t-i}\sigma_{t-j}$$
$$S_+ = \sum_{\tau=1}^{k} m_\tau \tag{4}$$
$$\sigma_Z^2 = \frac{\sum_{i,j=1}^{k}(m_i m_j \Sigma_{ij})}{S_+^2}$$
$$\mu_Z = \log S_+ - \frac{1}{2}\sigma_Z^2$$

With μ_Z and σ_Z defined, R_t is approximated as the ratio of two log-normals where $\Sigma_{0Z} = \text{Cov}(\log \lambda_t, \log S)$ is the covariance between the numerator and the log-denominator. Since $S = \sum_\tau X_\tau$, and using a first-order approximation, this covariance is a weighted average of the covariances between $\log \lambda_t$ and each $\log \lambda_{t-\tau}$, weighted by the relative expected contributions m_τ.:

$$R_t \sim \frac{Lognormal(\mu_t, \sigma_t)}{Lognormal(\mu_Z, \sigma_Z)}$$

$$\mu_{R_t} = \mu_t - \mu_Z = \mu_t - \log S_+ + \frac{1}{2}\sigma_Z^2$$

$$\sigma_{R_t} = \sqrt{\sigma_t^2 + \sigma_Z^2 - 2\Sigma_{0Z}} \tag{5}$$

$$\Sigma_{0Z} = \frac{\sum_{\tau=1}^{k} m_\tau \Sigma_{0\tau}}{\sum_{\tau=1}^{k} m_\tau} = \frac{1}{S_+} \sum_{\tau=1}^{k} m_\tau \Sigma_{0\tau}$$

$$R_t \sim Lognormal(\mu_{R_t}, \sigma_{R_t})$$

This formulation of R_t from (5) assumes knowledge of the posterior or prediction [70] covariance of the incidence estimates (Σ_{ij}). This is typical in modern [71] frameworks [20, 21], but in other situations may not be available. In this case we could [72] assume the individual estimates of the incidence are independent, however this alters [73] the uncertainty of our R_t estimate and in certain circumstances introduces a potential [74] underestimation bias, influenced by the true off-diagonal correlation mass and the [75] certainty of the incidence estimates. An alternative approach is to assume weak [76] stationarity and estimate a parametric correlation model from the data used to build the [77] incidence model, using Pearson residuals to parameterise an exponential decay function [78] based on time difference [22] (see supporting software package for implementation). The [79] degree of bias invovled in the assumption of independence is investigated further in S3 [80] Appendix, and heuristics for assessing the significance of this bias proposed. [81]

The method for estimating R_t from modelled incidence has been described assuming [82] a non-negative component to the infectivity profile, as it is implicit that infector and [83] infectee are necessarily sequential in time. In the situation where symptomatic case [84] counts are used as a proxy for incidence and the serial interval as a proxy for the [85] infectivity profile, negative times between serial cases may be observed due to variation [86] in delay in observation of the transmission chain. There is nothing in this framework to [87] stop the use of a negative time for the infectivity profile, and we can directly support R_t [88] estimates in these cases. [89]

Numerical stability [90]

In (5) μ_t is the log-scale mean of the incidence estimate at time t, and σ_t its standard [91] deviation. These can be large, leading to numerical instability in terms involving [92] $\exp(\mu + \sigma^2)$. However, assuming non-negative correlations and using log-space [93] computation with optimized log-sum-exp functions [23], the expressions remain [94] computationally tractable: [95]

$$\log m_\tau = \mu_{t-\tau} + \log \omega_\tau + \frac{1}{2}\sigma_{t-\tau}^2$$

$$\log S_+ = \text{logsumexp}_\tau(\log m_\tau)$$

$$\log \sigma_Z^2 = \text{logsumexp}_{i,j}(\log m_i + \log m_j + \log \Sigma_{ij}) - 2\log S_+ \tag{6}$$

$$\log \Sigma_{0Z} = \text{logsumexp}_\tau(\log m_\tau + \log \Sigma_{0\tau}) - \log S_+$$

Other relations may be implemented directly as in (5). [96]

Infectivity profile uncertainty [97]

This estimate of R_t is conditioned on a single known infectivity profile. In reality there [98] is also uncertainty in the infectivity profile (ω) which plays a role in the definition of [99]

$\mu_{Z,t}$ and $\sigma_{Z,t}$. We cannot assume any particular distributional form for the infectivity profile, but we can use a range of empirical estimates of the infectivity profile to calculate multiple distributional estimates for R_t and then combine these as a mixture distribution.

The nature of this mixture distribution will depend on the various estimates of the infectivity profile distributions. However, we can use general properties of mixture distributions to create estimates for the mean and variance of the reproduction number estimate (R_t^*) combining the uncertainty arising from multiple infection profile estimates (Ω) and from the incidence estimate model itself:

$$
\begin{aligned}
E[R_t|\omega] &= e^{(\mu_{R_t,\omega} - \frac{1}{2}\sigma^2_{R_t,\omega})} \\
V[R_t|\omega] &= \left[e^{(\sigma^2_{R_t,\omega})} - 1\right]\left[e^{2\mu_{R_t,\omega}+\sigma^2_{R_t,\omega}}\right] \\
E[R_t^*] &= \frac{1}{|\Omega|}\sum_{\omega\in\Omega} E[R_t|\omega] \\
V[R_t^*] &= \frac{1}{|\Omega|}\left[\sum_{\omega\in\Omega} V[R_t|\omega] + E[R_t|\omega]^2\right] - E[R_t^*]^2
\end{aligned}
\tag{7}
$$

The cumulative distribution function of the mixture is simply the arithmetic mean of the component cumulative distribution functions (conditioned on each infectivity profile). If Φ is the cumulative distribution function of the standard normal distribution:

$$
\begin{aligned}
F_{R_t^*}(x) &= \frac{1}{|\Omega|}\sum_{\omega\in\Omega} F_{R_t}(x|\omega) \\
P(R_t^* \le x) &= \frac{1}{|\Omega|}\sum_{\omega\in\Omega} P(R_{t,\omega} \le x) \\
P(R_t^* \le x) &= \frac{1}{|\Omega|}\sum_{\omega\in\Omega} \Phi\left(\frac{ln(x)-\mu_{R_t,\omega}}{\sigma_{R_t,\omega}}\right)
\end{aligned}
\tag{8}
$$

As the cumulative density function of this mixture distribution is a strictly increasing function, specific solutions for median ($q_{0.5}$) and 95% confidence intervals ($q_{0.025}$ and $q_{0.975}$) can be calculated numerically by solving the following equations:

$$
\begin{aligned}
\frac{1}{|\Omega|}\sum_{\omega\in\Omega} \Phi\left(\frac{ln(q_{0.025})-\mu_{R_t,\omega}}{\sigma_{R_t,\omega}}\right) - 0.025 &= 0 \\
\frac{1}{|\Omega|}\sum_{\omega\in\Omega} \Phi\left(\frac{ln(q_{0.5})-\mu_{R_t,\omega}}{\sigma_{R_t,\omega}}\right) - 0.5 &= 0 \\
\frac{1}{|\Omega|}\sum_{\omega\in\Omega} \Phi\left(\frac{ln(q_{0.975})-\mu_{R_t,\omega}}{\sigma_{R_t,\omega}}\right) - 0.975 &= 0
\end{aligned}
\tag{9}
$$

Numerical solutions to this are moderately expensive to perform. A reasonable approximation can be expected by matching moments of a log normal distribution to the mean $E[R_t^*]$ and variance $V[R_t^*]$ of the mixture. This gives us the final closed form estimator for the reproduction number given a set of infectivity profiles, $\overline{R_{t,\Omega}}$, as:

$$\mu_{t|\Omega} = log\left(\frac{E[R_t^*]^2}{\sqrt{E[R_t^*]^2 + V[R_t^*]}}\right)$$

$$\sigma_{t|\Omega} = \sqrt{log\left(1 + \frac{V[R_t^*]}{E[R_t^*]^2}\right)} \tag{10}$$

$$\overline{R_{t|\Omega}} \sim Lognormal(\mu_{t|\Omega}, \sigma_{t|\Omega})$$

In summary we present a method for retrieving the distributional form of the reproduction number from log normally distributed probabilistic estimates of incidence arising from statistical count models. This includes uncertainty arising from both count models and from infectivity profile distributions.

Validation

To test this method we developed a simulation based on a branching process model parametrised by 5 different R_t time series, a set number of imported infections at time zero, and fixed infectivity profile (see S1 Appendix Fig S1). Taken together, R_t and the infectivity profile define the expected number of secondary infections given a primary infection, on each day post infection. This expectation is sampled using a Poisson distribution to realise simulated infections on each day. In each simulation run, the degree of outward edges in the network of realised infections at any given time is an instantaneous R_t. For each paramterisation of R_t we generate 50 simulations with different random seeds, so that in bulk the simulation reproduction number will be close to the parameterised R_t. Each simulation generates a line list of synthetic infections. The line list of infected individuals were aggregated to daily counts of infection. Five random scenarios with different input R_t time series parametrisation were considered, and 50 replicates of each scenario were simulated, with different random seeds, resulting in 250 simulations.

We are particularly interested in uncertainty propagation. To assess the effect of noise in the input case counts in subsequent R_t estimates we assume that infection case counts are subject to varying degrees of ascertainment which change from day to day. The levels of ascertainment were applied to the same underlying infection time series, with observed counts being a binomial sample from the "true" infection counts for any given day. The probability of ascertainment on any given day was a random sample from a Beta distribution with a fixed mean, but three different coefficients of variation (parameter values in the S1 Appendix). In this way there are three versions of each of the 250 simulations which have the same underlying infection counts, but whose case counts only vary by the degree of statistical noise in the observation of infections.

The resulting 750 observed infection counts were used directly as an input to 'EpiEstim' to generate a set of baseline R_t estimates, using a window of 14 days. The synthetic infection time-series were also used as input to estimate the underlying infection rate in two ways. Firstly we used a simple statistical Poisson model with time varying rate parameter, represented by a piecewise polynomial of order 2, and fitted using maximum likelihood with a logarithmic link function according to the methods of Loader et al. [24] and a bandwidth equivalent to 14 days as implemented in the R package 'locfit' [25]. We used the central estimate and standard error of this as input to the 'R_t from incidence' algorithm, assuming indepence. We refer to this estimate as 'R_t Locfit'. Secondly we estimated the incidence rate using a more sophisticated generalised additive model (GAM) with a smooth time parameter, with regularly spaced knots, one per 14 days, as implemented by the package 'mgcv' [22]. We used the central estimates and full covariance matrix of the predictions as input to the 'R_t from incidence'

algortihm, and we refer to this estimate as 'R_t GAM'. The resulting sets of estimates of R_t are broadly equivalent to that produced by 'EpiEstim', with 'R_t Locfit' representing a simple implementation, and 'R_t GAM' a more sophisticated implementation of our algorithm.

All three estimators were analysed for estimation delays, by identifying the minimum root mean squared error between estimate and true values when applied to a synthetic dataset designed for this purpose. The lags were corrected for by shifting the R_t estimate by that appropriate number of days (see S1 Appendix Fig S2 and Table S1 for details).

For each estimator method and within the 3 groups of low, medium and high ascertainment noise, 5 scenarios were run with 50 replicates of each scenario. The posterior distributions of daily R_t estimates from both estimators, 'EpiEstim', 'R_t Locfit' and 'R_t GAM', were compared to simulation ground truth at all time points and summarised for each time series to give estimator performance metrics for each of the 750 simulation replicas. From the each of these replicas 20 bootstraps were resampled during summarisation (resulting in 15,000 sets of estimator metrics per method). Estimator metrics for each method were graphically summarised to the 3 groups and presented as box-plots (main paper) and summarised to each of 5 scenarios within each of the 3 groups (full details in S2 Appendix). Comparisons between methods were made graphically.

We calculated the average continuous ranked probability score (CRPS) as a overall performance metric [26–30]. The average proportional bias of the estimates within each simulation gives us a measure of estimator bias. We calculated the mean of the 50% interval width (inter-quartile range) of each estimate as a measure of estimator sharpness. We calculated the coverage probability of the 50% inter quantile range as an indicator of estimator calibration. We further investigated calibration by de-biasing estimates and with these adjusted estimates derived a novel calibration metric as the Wasserstein distance [31] of the probability integral transform histogram from the uniform [32–35] (lower values are better). To test the ability to discriminate between a growing or shrinking epidemic we calculated the prediction probability of misclassification of the true value R_t being greater than or less than 1. A weighted average of these by absolute distance of the true value from 1 gives us a estimator metric for this specific question (lower values are better). These metrics are fully defined in S2 Appendix.

We conducted two sensitivity analyses, one with an 'EpiEstim' window and 'locfit' bandwidth. or 'GAM' knots equivalent to 7 days, rather than the 14 in the main analysis, and a second comparing estimate quality when the estimates were not corrected for delays.

Use of large language models

We acknowledge the input of the large language model QWEN3-235B-A22B-2507, principally in refining the mathematical methods in this paper. All methods were conceptualised by the authors, but QWEN3 contributed to improve the handling of estimate covariance and assessment of the bias when assuming independence (S3 Appendix). QWEN3 was also used to improve mathematical notation and consistency. There was no use of LLMs in generating the final narrative of this paper, or in validation of the methods. The reference implementations of these methods was written without the use of LLMs, except for the code responsible for the generation of synthetic variance-covariance matrices. All mathematical and code generated by QWEN3 was rigourously reviewed and tested by the authors. Discussion history is available from the corresponding author on request.

Results

In Fig 1 panel A case counts and a modelled incidence estimate from a single simulation are shown for the 3 levels of ascertainment noise. Uncertainty in modelled incidence estimates increases with noise. In panel B we compare the R_t estimates derived from this this simulation. 'EpiEstim' can be observed to produce a slightly lagged estimate (top row panel B). In general the confidence of 'EpiEstim' estimates appear related to the distance of the estimate from 1. The central estimate becomes more volatile with more noise in the data set, but the confidence intervals do not appear to widen, suggesting noise in the input data does not affect uncertainty. By comparison 'R_t GAM' estimates (second row panel B) are smoother, less lagged and suffer from fewer tail effects at the limits of the time series. 'R_t Locfit' estimates (lower row panel B) also show no obvious lag, are overall more uncertain, particularly at the start and end of the time series. Increased noise in the data increases the uncertainty in the 'R_t Locfit' estimates more than the other two estimates.

In the main validation scenario we used a window for 'EpiEstim', 'GAM' and 'locfit' of 14 days. We saw in Fig 1 that this results in the estimate of R_t lagging the true value. This is quantified in Table 1 and Fig S2 in the S1 Appendix. In the main analysis 'EpiEstim' tends to produce an estimate delayed by 7 days and this lag is corrected by shifting the estimate in time before other metrics are calculated. In the first sensitivity analyses with a window of 7 days, a 4 day lag is observed for 'EpiEstim', and in the second sensitivity analysis the metrics are calculated without correcting for lags.

Table 1. Estimator delays in the validation scenarios

	Window (days)	
method	14	7
R_t Locfit	0	0
R_t GAM	0	0
EpiEstim	7	4

A quantification of the quality of the two estimation methods is shown in Fig 2 summarised from all 750 simulations, and corrected for estimator delays. In panel A the continuous ranked probability score is lower (better) for 'R_t GAM' and 'R_t Locfit' over all scenarios combined. In panel B the proportional difference between the true value and the median of the estimate probability (expressed as a percentage change) shows that the median of 'EpiEstim' tends to slightly overestimate the true R_t by around 2-3% whereas 'R_t Locfit' exhibits little bias, with 'R_t GAM' in between. In Fig 2 panel C, the 50% prediction interval width is smaller for 'EpiEstim' and 'R_t GAM' demonstrating sharper, more confident, estimates than the predictions of 'R_t Locfit'. 'R_t Locfit' is the only method which has evidence of decreasing sharpness (increasing interval width - Fig 2 panel C) with increasing data noise. 'EpiEstim' does not exhibit this change in sharpness with data noise and 'R_t GAM' only minimally. The low values of the probability of 50% prediction coverage, as seen in panel D, which is ideally 0.5, implies that 'EpiEstim' and 'R_t GAM' are over-confident in its predictions, 'R_t Locfit' on the other hand looks to be somewhat conservative. The PIT Wasserstein metric in Fig 2 panel E, allows us to compare the calibration of the 2 methods on the same scale, one of which is over confident and the other being under confident. This suggests 'R_t Locfit' is better calibrated than 'EpiEstim' or 'R_t GAM' (albeit in a different direction). Finally Fig 2 panel F shows the probability of misclassification at the threshold of $R_t = 1$ which is similar for 'EpiEstim' and 'R_t Locfit', but lower (better) for 'R_t GAM'. We also looked at how these metrics varied between scenarios (details in S1 Appendix Fig S3 and S4) and the breakdown largely follows the summary presented in Fig 2, although 'R_t GAM' performs relatively poorly in scenario 3, and in scenario 5, which
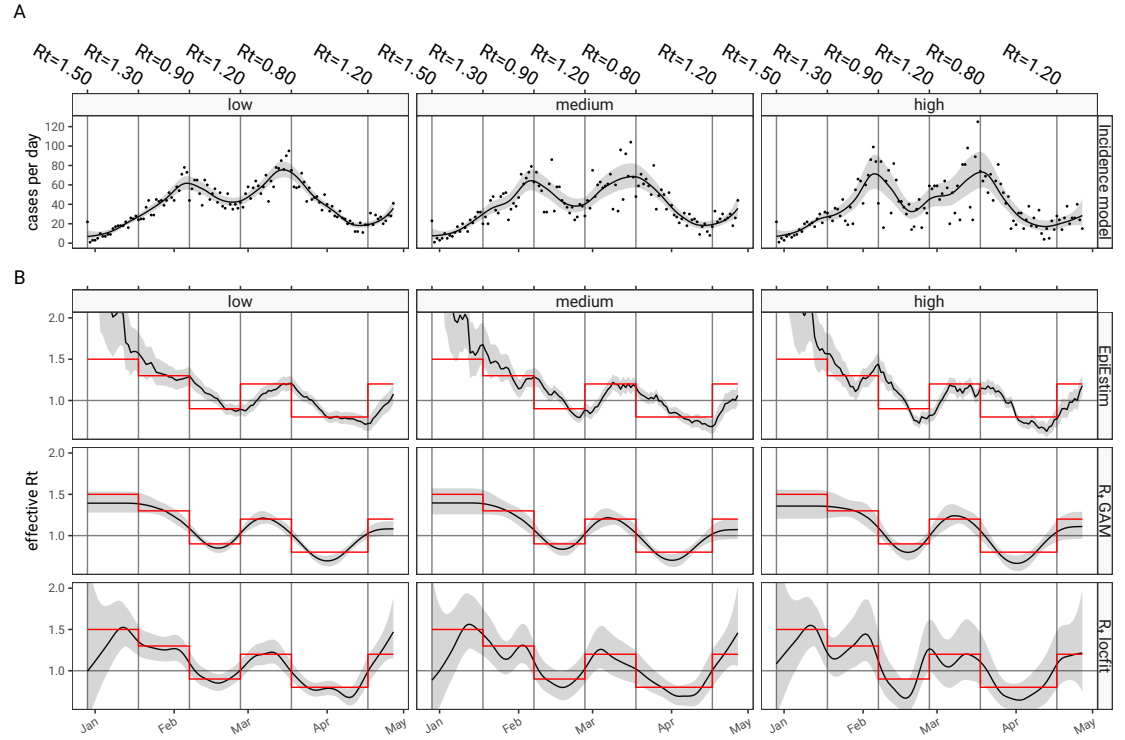
Fig 1. Instantaneous reproduction number estimates from a branching process model simulation. A qualitative comparison of instantaneous reproduction number estimates is shown. Panel A shows three case time series based on a single run of a branching process model parametrised with a stepped reproduction number time series (red lines in panel B) and infectivity profile as in S1 Appendix Fig S1. Case counts are shown as dots. A smoothed estimate of the cases per day as a line with shaded 95% confidence intervals, based on a simple Poisson regression model. All three time series have on average 70% case ascertainment, however the day to day variability of ascertainment is parameterised as a Beta distributed random variable, with "low", "medium" and "high" relating to the coefficient of variation of the Beta distribution (see S1 Appendix Table S1). Panel B shows estimates of the reproduction number based on the methods presented in this paper, and in the top row 'EpiEstim' estimates derived from the data points in panel A are shown. In the middle row R_t esimtates from a combination of GAM incidence model and the mthods described in the paper. In the bottom row, R_t estimates derived from a 'locfit' incidence model. In panel B the parameterised R_t is shown as a solid red line and can be regarded as the ground truth for this single simulation run.

has relatively little variation compared to the others, calibration is better for 'EpiEstim', and 'R_t GAM' appears overconfident.

In the first sensitivity analysis (S1 Appendix Fig S5), the amount of data informing the R_t estimates is reduced from 14 to 7 days, resulting in reduced certainty in both estimators. This changes the relative performance of the 2 methods as measured the the CRPS, which is now tends to favour 'R_t GAM' as the better overall estimator, despite continued bias, excess sharpness and mis-calibration. The certainty of 'R_t Locfit' has dropped to a low level and it has become excessively conservative in all but the high ascertainment noise scenarios (further details in S1 Appendix). This also slightly increases the probability of misclassification at the threshold of $R_t = 1$ for 'R_t Locfit',
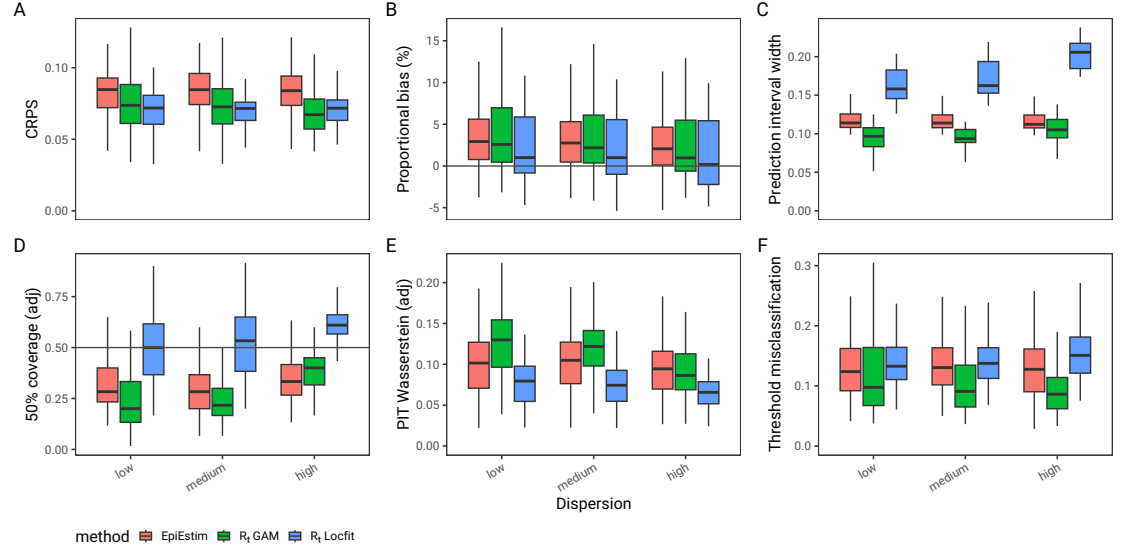
Fig 2. Quantitative comparison of R_t estimation methods applied to 50 simulations of 5 scenarios at 3 levels of ascertainment noise. The figure compares metrics describing the overall performance of the estimators. In Panel A is the continuous ranked probability score (CRPS) - lower is better; the average proportional bias (panel B) which characterise bias - lower is better; In panel C, the 50% prediction interval width measures estimator sharpness, and lower is better if the estimator is unbiased and well calibrated. In Panel D the probability of 50% coverage (ideal value is 0.5), and in Panel E the adjusted probability integral transform (PIT) Wasserstein metric (lower is better) are both measures of calibration. In Panel F a functional metric which quantifies the probability that an estimate is the wrong side of the circical threshold of $R_t = 1$ desribes utility in decision making.

although 'R_t GAM' still outperforms other methods on this more functional metric. In the second sensitivity analysis with no correction for lag (details in S1 Appendix Fig S6), 'EpiEstim', which is affected by lag, performs much less well in all metrics as the lag penalises all dimensions of the quality of the estimator, and this serves to highlight the importance of addressing lag before comparing the bias and calibration of estimators.

Discussion

In this paper we describe a mathematical method for deriving an estimate of the effective reproduction number (R_t) from modelled estimates of disease incidence, that incorporates uncertainty from both incidence model and infectivity profile. We provide evidence that when combined with incidence models, our method produces R_t estimates that are comparable to the de-facto standard algorithm implemented in 'EpiEstim' [6], when assessed against synthetic outbreak data. Depending on the nature of the incidence estimate used as input our R_t estimation can reflect uncertainty in the data.

There is a similar approach to deriving R_t from modelled incidence described by Gressani et al. [11]. It assumes a specific formulation for the incidence model as a negative binomial distribution estimated with P-splines in a Bayesian framework with Laplace approximations. This encouragingly also gives log normally distributed estimates for R_t, but its methodology is predicated on the specifics of the P-spline posterior vector and it does not incorporate infectivity profile uncertainty.

By comparison the approach in this paper is only loosely coupled to the incidence estimate framework and so can be applied to any incidence model that produces a time varying log-normally distributed incidence estimate. This includes a broad family of Poisson and negative binomial regression models [20, 24, 36], or latent Gaussian models [21] using logarithmic link functions, and is agnostic to the formulation of those models. The incidence models can be estimated on time aggregated data [9], include covariates such as day of week effects, incorporate change points without affecting the derivation of the reproduction number estimates.

In terms of infectivity profile, our method is robust to distributions with zero or negative time intervals between index and secondary observations. It could therefore be used with directly observed real world serial interval distributions [18] and delayed case counts as proxies for infectivity profile and infection incidence, which are necessarily inferred quantities; this is explored further in the 'ggoutbreak' package documentation (`https://ai4ci.github.io/ggoutbreak/`). Our method does not specifically address important questions arising from ascertainment bias, or right truncation of observed incidence [3, 16], however when input indidence models have been adapted to right censored data our method can be used to derive an R_t estimate (see package documentation for more details).

The validation comparison here combines one very simple statistical model of incidence 'R_t Locfit', and one more sophisticated model 'R_t GAM', with our method to derive R_t estimates, and compares them to estimates produced direct from the count data by 'EpiEstim'. This shows that the combination of incidence models and our R_t derivation, produces estimates similar to 'EpiEstim'. We have not formally tested the statistical significance of the differences because they are based on a large number of observations, and even tiny differences will be statistically significant. We pragmatically chose to simulate using a set of 5 step functions for R_t parametrisation, which we expect to be relatively challenging for both 'EpiEstim' and the statistical models for incidence we used here. We could have picked any other function to use for validation. It is clear that relative performance between the methods varies with the exact details of the test scenario (see S1 Appendix Fig S6). A more realistic smooth R_t parametrisation time series has been performed and both methods perform better in such scenarios, so we consider our simulations to be worst case. There are some scenarios in which 'EpiEstim' performs better and others in which incidence model derived estimates performs better. In comparing the models we ignored the first 20 R_t estimates which are unstable in 'EpiEstim' and very uncertain in 'R_t Locfit', this will tend to discriminate against 'R_t GAM' which seems comparatively accurate in the first 20 days.

Our method is not a replacement for 'EpiEstim' as it requires an incidence model dervied from count data, rather than directly using data, and any comparison is looking at both the incidence model quality and the method for deriving R_t. This must be bourne in mind when interpreting the validation section of this paper. By picking two statistical models for incidence with different characteristics, to which to apply our method for estimating R_t, we hope to demonstrate that the combinations are not obviously inferior to 'EpiEstim'. If we had used different incidence models, the overall R_t estimate may have very different characteristics. The choice of best estimator is subjective as it depends whether it is more important to have an estimator that minimises the risk of a misclassification a growing or shrinking epidemic, or whether a more honest representation of uncertainty is required. Estimates of later time points are also arguably more important in managing an epidemic.

Not withstanding these limitations in validation, we argue our method for deriving R_t from log-normally modelled incidence estimates, which are commonly output by statistical modelling frameworks, is a useful adjunct to the range of tools available for monitoring an epidemic. It is relatively quick and deterministic, and is flexible enough

to be combined with a wide range of temporal incidence modelling techniques, which 336
can account for reporting delays or ascertainment bias, or could be extended to 337
spatio-temporal incidence models. 338

References

1. Gostic KM, McGough L, Baskerville EB, Abbott S, Joshi K, Tedijanto C, et al. Practical Considerations for Measuring the Effective Reproductive Number, Rt;16(12):e1008409. doi:10.1371/journal.pcbi.1008409.

2. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, et al. Estimated Transmissibility and Impact of SARS-CoV-2 Lineage B.1.1.7 in England;372(6538). doi:10.1126/science.abg3055.

3. Abbott S, Hellewell J, Sherratt K, Gostic K, Hickson J, Badr HS, et al.. EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters;. Available from: `https://epiforecasts.io/EpiNow2/`.

4. Alvarez L, Colom M, Morel JD, Morel JM. Computing the Daily Reproduction Number of COVID-19 by Inverting the Renewal Equation Using a Variational Technique;118(50):e2105112118. doi:10.1073/pnas.2105112118.

5. Parag KV. Improved Estimation of Time-Varying Reproduction Numbers at Low Case Incidence and between Epidemic Waves;17(9):e1009347. doi:10.1371/journal.pcbi.1009347.

6. Thompson RN, Stockwin JE, van Gaalen RD, Polonsky JA, Kamvar ZN, Demarsh PA, et al. Improved Inference of Time-Varying Reproduction Numbers during Infectious Disease Outbreaks;29:100356. doi:10.1016/j.epidem.2019.100356.

7. Wallinga J, Lipsitch M. How Generation Intervals Shape the Relationship between Growth Rates and Reproductive Numbers;doi:10.1098/rspb.2006.3754.

8. Steyn N, Parag KV. Robust Uncertainty Quantification in Popular Estimators of the Instantaneous Reproduction Number;. Available from: `https://www.medrxiv.org/content/10.1101/2024.10.22.24315918v1`.

9. Nash RK, Bhatt S, Cori A, Nouvellet P. Estimating the Epidemic Reproduction Number from Temporally Aggregated Incidence Data: A Statistical Modelling Approach and Software Tool;19(8):e1011439. doi:10.1371/journal.pcbi.1011439.

10. Nash RK, Nouvellet P, Cori A. Real-Time Estimation of the Epidemic Reproduction Number: Scoping Review of the Applications and Challenges;1(6):e0000052. doi:10.1371/journal.pdig.0000052.

11. 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(10):e1010618. doi:10.1371/journal.pcbi.1010618.

12. Cauchemez S, Boëlle PY, Thomas G, Valleron AJ. Estimating in Real Time the Efficacy of Measures to Control Emerging Communicable Diseases;164(6):591–597. doi:10.1093/aje/kwj274.

13. Hong HG, Li Y. Estimation of Time-Varying Reproduction Numbers Underlying Epidemiological Processes: A New Statistical Tool for the COVID-19 Pandemic;15(7):e0236464. doi:10.1371/journal.pone.0236464.

14. Johnson KD, Beiglböck M, Eder M, Grass A, Hermisson J, Pammer G, et al. Disease Momentum: Estimating the Reproduction Number in the Presence of Superspreading;6:706–728. doi:10.1016/j.idm.2021.03.006.

15. Ogi-Gittins I, Hart WS, Song J, Nash RK, Polonsky J, Cori A, et al. A Simulation-Based Approach for Estimating the Time-Dependent Reproduction Number from Temporally Aggregated Disease Incidence Time Series Data;47:100773. doi:10.1016/j.epidem.2024.100773.

16. Abbott S, Hellewell J, Thompson RN. Estimating the Time-Varying Reproduction Number of SARS-CoV-2 Using National and Subnational Case Counts [Version 2; Peer Review: 1 Approved, 1 Approved with Reservations];5(112). doi:10.12688/wellcomeopenres.16006.2.

17. Vegvari C, Abbott S, Ball F, Brooks-Pollock E, Challen R, Collyer BS, et al. Commentary on the Use of the Reproduction Number R during the COVID-19 Pandemic;31(9):1675–1685. doi:10.1177/09622802211037079.

18. Park SW, Sun K, Champredon D, Li M, Bolker BM, Earn DJD, et al. Forward-Looking Serial Intervals Correctly Link Epidemic Growth to Reproduction Numbers;118(2):e2011548118. doi:10.1073/pnas.2011548118.

19. Lo CF. WKB Approximation for the Sum of Two Correlated Lognormal Random Variables;7(129):6355–6367. doi:10.2139/ssrn.2220803.

20. Hastie TJ. Generalized Additive Models. Routledge;.

21. Rue H, Martino S, Chopin N. Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations;71(2):319–392. doi:10.1111/j.1467-9868.2008.00700.x.

22. Wood SN. Generalized Additive Models: An Introduction with R, Second Edition. 2nd ed. New York: Chapman and Hall/CRC; 2017.

23. Blanchard P, Higham DJ, Higham NJ. Accurately Computing the Log-Sum-Exp and Softmax Functions;41(4):2311–2330. doi:10.1093/imanum/draa038.

24. Loader C. Local Regression and Likelihood. Statistics and Computing. Springer-Verlag;. Available from: http://link.springer.com/10.1007/b98858.

25. Loader C, Sun J, Technologies L, Liaw A. Locfit: Local Regression, Likelihood and Density Estimation;. Available from: https://CRAN.R-project.org/package=locfit.

26. Anderson JL. A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations;9(7):1518–1530. doi:10.1175/1520-0442(1996)009¡1518:AMFPAE¿2.0.CO;2.

27. Bosse NI, Gruson H, Cori A, van Leeuwen E, Funk S, Abbott S. Evaluating Forecasts with Scoringutils in R;. Available from: http://arxiv.org/abs/2205.07090.

28. Bosse NI, Abbott S, Cori A, van Leeuwen E, Bracher J, Funk S. Scoring Epidemiological Forecasts on Transformed Scales;19(8):e1011393. doi:10.1371/journal.pcbi.1011393.

29. Bröcker J. On Reliability Analysis of Multi-Categorical Forecasts;15(4):661–673. doi:10.5194/npg-15-661-2008.

30. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation;102(477):359–378. doi:10.1198/016214506000001437.

31. Panaretos VM, Zemel Y. Statistical Aspects of Wasserstein Distances;6:405–431. doi:10.1146/annurev-statistics-030718-104938.

32. David FN, Johnson NL. The Probability Integral Transformation When Parameters Are Estimated from the Sample;35(1/2):182–190. doi:10.2307/2332638.

33. Hamill TM. Interpretation of Rank Histograms for Verifying Ensemble Forecasts;129(3):550–560. doi:10.1175/1520-0493(2001)129¡0550:IORHFV¿2.0.CO;2.

34. Wilks DS. Indices of Rank Histogram Flatness and Their Sampling Properties;147(2):763–769. doi:10.1175/MWR-D-18-0369.1.

35. Brockwell AE. Universal Residuals: A Multivariate Transformation;77(14):1473–1478. doi:10.1016/j.spl.2007.02.008.

36. Nelder JA, Wedderburn RWM. Generalized Linear Models;135(3):370–384. doi:10.2307/2344614.

Funding

Competing interests

The authors have no competing interests to declare.

Author contributions

RC and LD generated the research questions. RC performed the mathematical analysis and simulations, and created the supporting software package. RC and LD provided validation of the methods. LD provided supervision of the research. RC developed the first draft of the manuscript. RC and LD contributed to the final editing of the manuscript and its revision for publication and had responsibility for the decision to publish.

Data and code availability

All data and code used for running experiments, model fitting, and plotting is available on a GitHub repository at `https://ai4ci.github.io/ggoutbreak-paper/`. The methods described here are implemented in the form of an R package to support the

estimation of epidemiological parameters and it is deployed on the A14CI r-universe (`https://ai4ci.r-universe.dev/ggoutbreak`). We have also used Zenodo to assign a DOI to the repository: doi:10.5281/zenodo.7691196.

Supporting information

S1 Appendix. Simulation for validating Rt estimates, additional results and sensitivity analyses Methodological details of simulation set up for validating the performance of R_t estimators presented in the main paper, additional figures and detailed results from the sensitivity analyses.

S2 Appendix. Metrics for evaluating the quality of probabilistic estimators. Methodological details of performance metrics used for validating the performance of R_t estimators presented in the main paper.

S3 Appendix. Bias in R_t Estimation Under Independence and Heuristics for Risk Assessment. This appendix provides bounding of bias in R_t estimates under the assumption of independence of incidence estimates, and derives of metrics to assess whether the assumption of independence is valid.