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

TODO

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 [3–14], 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 [15].

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. [16], 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 when secondary infections occur with respect to primary infections. This is described as the infectivity profile, a time dependent probability distribution, and is equivalent to the generation time distribution [1, 17]. 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 [6,17]. 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, which we from this point refer to as 'R_t from incidence'. We validate this 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):

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

giving us:

$$\begin{aligned} R_t &\sim \frac{\lambda_t}{\sum_\tau \omega_\tau \lambda_{t-\tau}} \\ R_t &\sim \frac{Lognormal(\mu_t, \sigma_t)}{\sum_\tau Lognormal(\mu_{t-\tau} + log(\omega_\tau), \sigma_{t-\tau})} \end{aligned} \tag{2}$$

As an aside, it has been shown that the sum of log-normal distributions can be approximated by another log-normal [18] with parameters μ_Z and σ_Z.

$$\begin{aligned} S_+ &= \mathrm{E}\left[\sum_i X_i\right] = \sum_i \mathrm{E}[X_i] \\ &= \sum_i e^{\mu_i + \frac{1}{2}\sigma_i^2} \\ \sigma_Z^2 &= \frac{1}{S_+^2} \sum_{i,j} \mathrm{cor}_{ij}\, \sigma_i \sigma_j\, \mathrm{E}[X_i]\, \mathrm{E}[X_j] \\ &= \frac{1}{S_+^2} \sum_{i,j} \mathrm{cor}_{ij}\, \sigma_i \sigma_j e^{\mu_i + \frac{1}{2}\sigma_i^2} e^{\mu_j + \frac{1}{2}\sigma_j^2} \\ \mu_Z &= \ln(S_+) - \frac{1}{2}\sigma_Z^2 \end{aligned} \tag{3}$$

The sum term in the denominator of the renewal equation consists of a set of correlated scaled log normal distributions with both scale and correlation defined by the infectivity profile (ω). In our case cor_{ij} can be equated to the infectivity profile ($\omega_{|i-j|}$) when $i \neq j$ and 1 when $i = j$. μ_i is $\mu_{t-\tau} + ln(\omega_\tau)$.

$$
\begin{aligned}
S_t &= \sum_{s=1}^{|\omega|} \omega_s e^{\mu_{t-s} + \frac{1}{2}\sigma^2_{t-s}} \\
\sigma_{Z,t} &= \sqrt{\frac{\sum_{i,j=1}^{|\omega|} \left[\omega_{|i-j|} + I(i,j)\right] \omega_i \omega_j \left[\sigma_{(t-i)} e^{\mu_{(t-i)} + \frac{1}{2}\sigma^2_{(t-i)}}\right] \left[\sigma_{(t-j)} e^{\mu_{(t-j)} + \frac{1}{2}\sigma^2_{(t-j)}}\right]}{S_t^2}} \\
\mu_{Z,t} &= \log\left(S_t\right) - \frac{1}{2}\sigma^2_{Z,t}
\end{aligned}
\tag{4}
$$

μ is the central estimate of case counts on the log scale, and its standard deviation σ can be large. There are numerical stability issues dealing with terms involving $e^{(\mu+\sigma^2)}$, however keeping everything in log space and using optimised log-sum-exp functions this can be made computationally tractable [19].

$$
\begin{aligned}
\log(S_t) &= \log(\sum_{s=1}^{|\omega|} e^{\mu_{t-s} + \frac{1}{2}\sigma^2_{t-s} + \log(\omega_s)}) \\
\log(T_{t,\tau}) &= \log(\omega_\tau) + \log(\sigma_{(t-\tau)}) + \mu_{(t-\tau)} + \frac{1}{2}\sigma^2_{(t-\tau)}) \\
\log(cor_{i,j}) &= \log(\omega_{|i-j|} + I(i=j)) \\
\log(\sigma^2_{Z,t}) &= \log(\sum_{i,j=1}^{|\omega|} e^{\log(cor_{i,j}) + \log(T_{t,i}) + \log(T_{t,j})}) - 2\log(S_t) \\
\mu_{Z,t} &= \log(S_t) - \frac{1}{2}\sigma^2_{Z,t}
\end{aligned}
\tag{5}
$$

N.B. if we assume the individual estimates of the incidence are uncorrelated this simplifies to:

$$
\log(\sigma^2_{Z,t}) = \log(\sum_{\tau=1}^{|\omega|} e^{2\log(T_{t,\tau})}) - 2\log(S_t) \tag{6}
$$

With $\mu_{Z,t}$ and $\sigma_{Z,t}$ we can derive a distributional form of R_t incorporating uncertainty from modelled incidence estimates:

$$
\begin{aligned}
R_t &\sim \frac{Lognormal(\mu_t, \sigma_t)}{Lognormal(\mu_{Z,t}, \sigma_{Z,t})} \\
\mu_{R_t} &= \mu_t - \mu_{Z,t} \\
\sigma_{R_t} &= \sqrt{\sigma_t^2 + \sigma_{z,t}^2} \\
R_t &\sim Lognormal(\mu_{R_t}, \sigma_{R_t})
\end{aligned}
\tag{7}
$$

This estimate of R_t is conditioned on a single known infectivity profile. In reality there is also uncertainty in the infectivity profile (ω) which plays a role in the definition of $\mu_{Z,t}$ 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 the various empirical 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:

$$\begin{aligned}
E[R_t|\omega] &= e^{(\mu_{R_t,\omega} - \frac{1}{2}\sigma^2_{R_t,\omega})} \\
V[R_t|\omega] &= \left[e^{(\sigma^2_{R_t,\omega})} - 1\right]\left[e^{2\mu_{R_t,\omega} + \sigma^2_{R_t,\omega}}\right] \\
E[R_t^*] &= \frac{1}{|\Omega|}\sum_{\omega\in\Omega} E[R_t|\omega] \\
V[R_t^*] &= \frac{1}{|\Omega|}\left[\sum_{\omega\in\Omega} V[R_t|\omega] + E[R_t|\omega]^2\right] - E[R_t^*]^2
\end{aligned} \tag{8}$$

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:

$$\begin{aligned}
F_{R_t^*}(x) &= \frac{1}{|\Omega|}\sum_{\omega\in\Omega} F_{R_t}(x|\omega) \\
P(R_t^* \leq x) &= \frac{1}{|\Omega|}\sum_{\omega\in\Omega} P(R_{t,\omega} \leq x) \\
P(R_t^* \leq x) &= \frac{1}{|\Omega|}\sum_{\omega\in\Omega} \Phi\left(\frac{ln(x) - \mu_{R_t,\omega}}{\sigma_{R_t,\omega}}\right)
\end{aligned} \tag{9}$$

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:

$$\begin{aligned}
\frac{1}{|\Omega|}\sum_{\omega\in\Omega} \Phi\left(\frac{ln(q_{0.025}) - \mu_{R_t,\omega}}{\sigma_{R_t,\omega}}\right) - 0.025 = 0 \\
\frac{1}{|\Omega|}\sum_{\omega\in\Omega} \Phi\left(\frac{ln(q_{0.5}) - \mu_{R_t,\omega}}{\sigma_{R_t,\omega}}\right) - 0.5 = 0 \\
\frac{1}{|\Omega|}\sum_{\omega\in\Omega} \Phi\left(\frac{ln(q_{0.975}) - \mu_{R_t,\omega}}{\sigma_{R_t,\omega}}\right) - 0.975 = 0
\end{aligned} \tag{10}$$

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:

$$
\begin{aligned}
\mu_{t|\Omega} &= log\left(\frac{E[R_t^*]^2}{\sqrt{E[R_t^*]^2 + V[R_t^*]}}\right) \\
\sigma_{t|\Omega} &= \sqrt{log\left(1 + \frac{V[R_t^*]}{E[R_t^*]^2}\right)} \\
\overline{R_{t|\Omega}} &\sim Lognormal(\mu_{t|\Omega}, \sigma_{t|\Omega})
\end{aligned}
\tag{11}
$$

In summary we present a method for retrieving the distributional form of the reproduction number from estimates of incidence arising from simple statistical count models. This includes uncertainty arising from both count models and from infectivity profile distributions. It is fully deterministic and computationally inexpensive. It does not place any particular constraints on the nature of the infectivity profile distribution and can handle distributions that have a negative component, as is sometimes seen in serial interval estimates.

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 Supplementary 1 Fig 1 for details). 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 supplementary materials). 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, using 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. [20] and a bandwidth equivalent to 14 days as implemented in the R package 'locfit' [21]. The central estimate and the standard error of the fitted Poisson rate from this simple statical incidence model were used as inputs to the R_t estimation method described in

this paper, to derive an set of estimates of R_t which are broadly equivalent to that produced by 'EpiEstim'. This combined estimator is referred to as 'R_t from incidence'.

Both 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 supplementary Fig 2 and Table 1 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' and 'R_t from incidence', 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 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 (supplementary). Comparisons between the two methods are made graphically.

We calculated the average continuous ranked probability score (CRPS) as a overall performance metric [22–26]. 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 [27] of the probability integral transform histogram from the uniform [28–31]. 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. These metrics are more fully described in the supplementary materials.

We conducted two sensitivity analyses, one with an 'EpiEstim' window and 'locfit' bandwidth 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.

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 change, suggesting noise in the input data does not affect uncertainty. Comparing to this 'R_t from incidence' (lower row panel B) is subject to less lag, and overall more uncertain, but otherwise agrees well with the 'EpiEstim' estimate. Increased noise in the data increases the uncertainty of the Poisson rate estimates (panel A) and hence 'R_t from incidence' estimates, unlike 'EpiEstim'.

In the main validation scenario we used a window for 'EpiEstim' 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 Supplementary 1. In the main analysis 'EpiEstim' tends to produce an estimate delayed by 7 days and this lag is corrected by

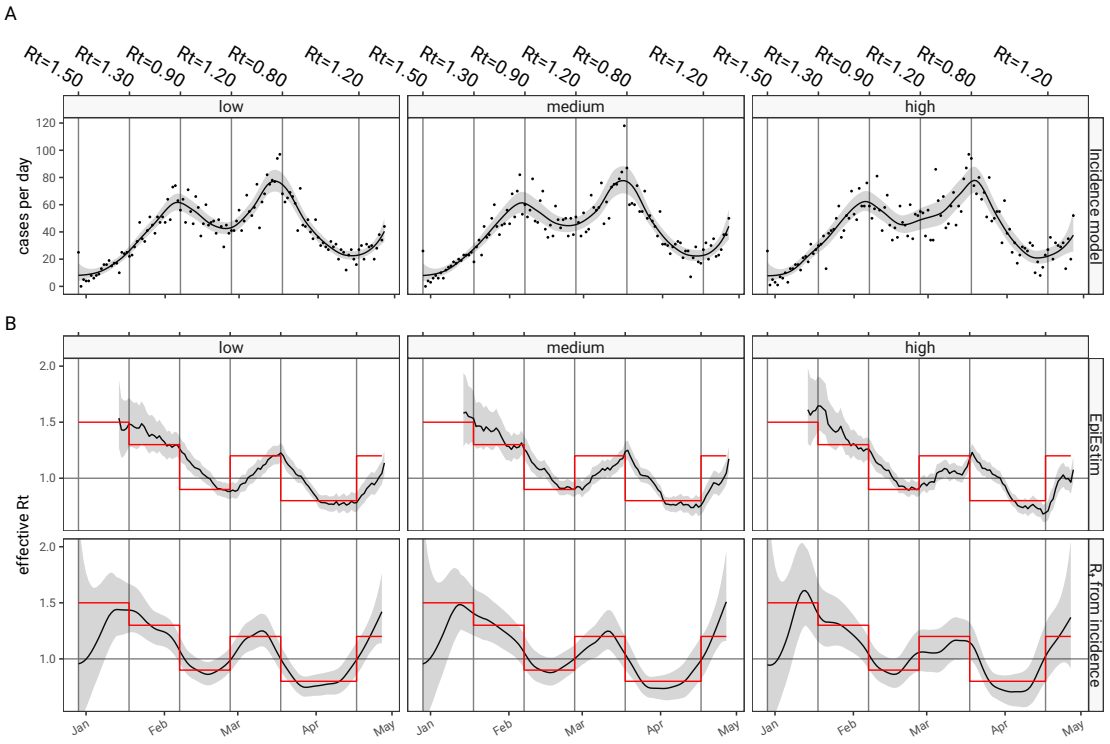

Fig 1. Instantaneous reproduction number estimates from a branching process model simulation. A qualitative comparison of instantaneous reproduction number estimates is shown. Panel A shows three case time series based on a single run of a branching process model parametrised with a stepped reproduction number time series (red lines in panel B) and infectivity profile as in supplementary figure 1. 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 supplementary materials). 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 bottom row, R_t estimates derived from the incidence model in Panel A using the methods described in this paper are shown. 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.

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 from incidence	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 from incidence' 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 from incidence' exhibits little bias. In Fig 2 panel C, the 50% prediction interval width is smaller for 'EpiEstim' demonstrating a sharper, more confident, estimate than the predictions of 'R_t from incidence', which has evidence of appropriately decreasing sharpness (increasing interval width - Fig 2 panel C) with increasing data noise. 'EpiEstim' does not exhibit this change in sharpness with data noise. The low values of the probability of 50% prediction coverage, as seen in panel D, which is ideally 0.5, implies that 'EpiEstim' is over-confident in its predictions, 'R_t from incidence' 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 marginally underconfident. This suggests 'R_t from incidence' is better calibrated than 'EpiEstim'. The final panel Fig 2 panel F shows the probability of misclassification at the threshold of $R_t = 1$ which is similar for both methods. We also looked at how these metrics varied between scenarios (details in supplementary 1 Fig 3 and 4) and the breakdown largely follows the summary presented in Fig 2, with the caveat that in scenario 5, which has relatively little variation compared to the others, calibration is better for 'EpiEstim'.

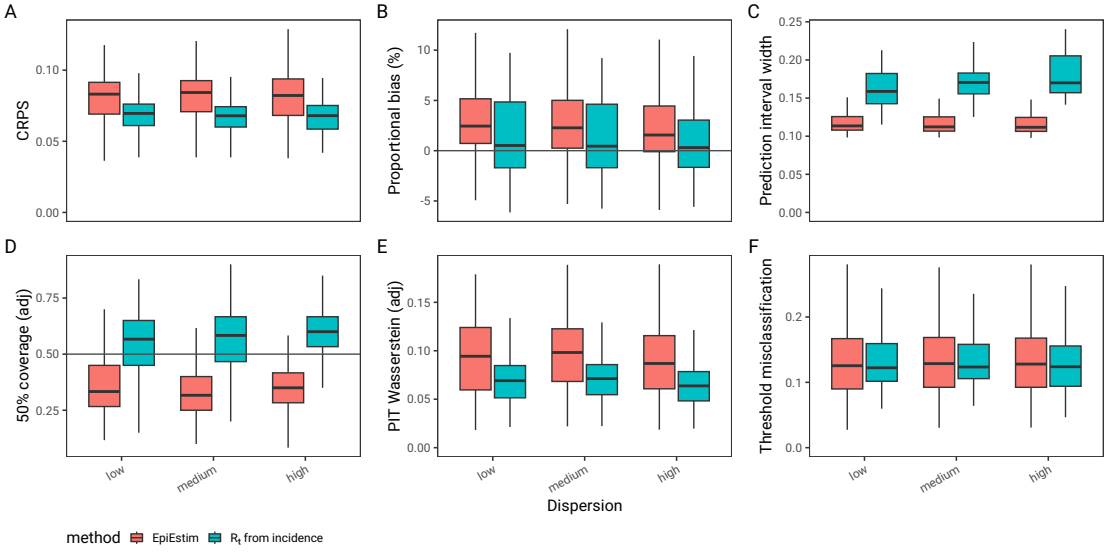

Fig 2. Quantitative comparison of R_t estimation methods applied to 50 simulations of 5 scenarios at 3 levels of ascertainment noise. The figure compares metrics describing the overall performance of the estimators with the continuous ranked probability score (CRPS) - lower is better, panel A; the average proportional bias, and mean universal residuals which characterise bias, in panel B and C - lower is better; The 50% prediction interval width measures estimator sharpness and lower is better if the estimator is unbiased and well calibrated (panel D); and the probability of 50% coverage (ideal value is 0.5, panel E), and adjusted probability integral transform (PIT) Wasserstein metric (lower is better, panel F)

In the first sensitivity analysis (Supplementary 1 Fig 5), 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 'EpiEstim' as the better overall estimator, despite continued bias, excess sharpness and miscalibration , partly because the certainty of 'R_t from incidence' has dropped to a low level and it has become excessively conservative in all but the high ascertainment noise scenarios (further details in the supplementary). This also slightly increases the probability of misclassification at the threshold of $R_t = 1$ for 'R_t from incidence'. In the second sensitivity analysis with no correction for lag (details in supplementary 1 Fig 6), '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 highlights 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 simple 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. In contrast to 'EpiEstim' our method produces estimates that reflect uncertainty in the data.

The method described here can be applied to any incidence model that produces time varying log-normally distributed incidence estimate. This includes a broad family of Poisson regression models [20, 32, 33], or latent Gaussian models [34] 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, or 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 [17] 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/`). However, our method does not specifically address important questions arising from ascertainment bias, or right truncation of observed incidence, both of which would have to be addressed in the pre-requisite incidence models [3, 15].

The validation comparison here uses a simple statistical model of incidence, combined with our method to derive R_t estimates, versus estimates produced direct from the data by 'EpiEstim'. This shows that the simple incidence model, and our R_t derivation, produces estimates that are not lagged, with slightly less bias than 'EpiEstim', and which have a more appropriate degree of confidence given their accuracy. We have not formally tested the statistical significance of these differences because they are based on a very large number of observations, and eveny 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 simple statistical models for incidence we used here. We could have picked any other function to use for validation, and it is clear that relative performance between the methods varies with the exact details of the test scenario (see supplementary materials). A more realistic smooth R_t paramterisation 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 'R_t from incidence' performs better.

Our method is not a like for like replacement for 'EpiEstim' as it requires an

incidence model, rather than directly using data, and comparing the two algorithms is therefore looking at both the incidence model and the method for deriving R_t. This must be bourne in mind when interpreting the validation section of this paper. By picking a very simple statistical model for incidence, to which to apply our method for estimating R_t, we hope to demonstrate that the combination is not obviously inferior to 'EpiEstim'. If we had used more sophisticated incidence model the overall R_t estimate may have very different characteristics, and it is clear that the configuration of the estimators affects the performance of the estimators.

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 could account for reporting delays or ascertainment bias, or extended to spatio-temporal incidence models.

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, family=Gaalen pu given=R D, 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. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

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

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

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

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

22. 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.

23. 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`.

24. 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.

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

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

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

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

29. 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.

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

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

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

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

34. 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.

Funding

RC and LD are funded by UK Research and Innovation AI programme of the Engineering and Physical Sciences Research Council (EPSRC grant EP/Y028392/1; `https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/Y028392/1`). RC and LD are affiliated with the JUNIPER partnership funded by the Medical Research Council (MRC grant MR/X018598/1; `https://www.ukri.org/councils/mrc/`). The views expressed are those of the authors.

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.