**Supplementary Material for *A simulation-based approach for estimating the time-dependent reproduction number from temporally aggregated disease incidence time series data***

I Ogi-Gittins1,2, WS Hart3, J Song4, RK Nash5, J Polonsky6, A Cori5, EM Hill1,2, RN Thompson1,2,3

**Affiliations:**

1Mathematics Institute, University of Warwick, Coventry, CV4 7AL, UK

2Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research (SBIDER), University of Warwick, Coventry, CV4 7AL, UK

3Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK

4Communicable Disease Surveillance Centre, Health Protection Division, Public Health Wales, Swansea, SA2 8QA, UK

5MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College, London, W2 1PG, UK

6Geneva Centre of Humanitarian Studies, University of Geneva, Geneva, 1205, Switzerland

**Supplementary Text**

**Discretisation of the serial interval**

Here, we explain how a continuous serial interval distribution with probability density function can be discretised into timesteps of length weeks to obtain ( and , both of which are defined in the main text.We adapt the approach described by Cori *et al.* [1] (see web appendix 11 of that article) in which the serial interval is discretised into timesteps of length .

We consider an infector-infectee transmission pair, with the infectee arising in the disease incidence time series data timesteps after their infector, where each timestep is of length weeks. Assuming that both the infector and infectee develop symptoms at a time that is uniformly distributed within those timesteps, then the probability density function of the period between their symptom onset times is

To discretise the serial interval distribution, the probability density function above is then weighted with the probability density function of the continuous serial interval distribution, to give

In principle, the calculation above can be applied when , and a similar argument can be used to obtain the probability of an infector and infectee appearing in the disease incidence time series in the same timestep (which would correspond to ). However, since the renewal equation model requires all new cases in a given timestep to have been infected by infectors appearing in the incidence data at a strictly earlier timestep, rather than the same timestep, we neglect and instead assume that same timestep infections are assigned to . In other words, we simply set so that sums to one.

When we apply the Cori method, we require the continuous serial interval distribution to be discretised into weekly timesteps. This therefore corresponds to undertaking the above calculations with .

**Simulation-based inference of**

Here, we give further details about the simulation-based method. The value of (for ) is estimated iteratively: in other words, is estimated first, followed by , and so on. By estimating iteratively, our inference procedure can be performed more quickly than attempting to estimate for all values of simultaneously (as in standard ABC rejection sampling [2]).

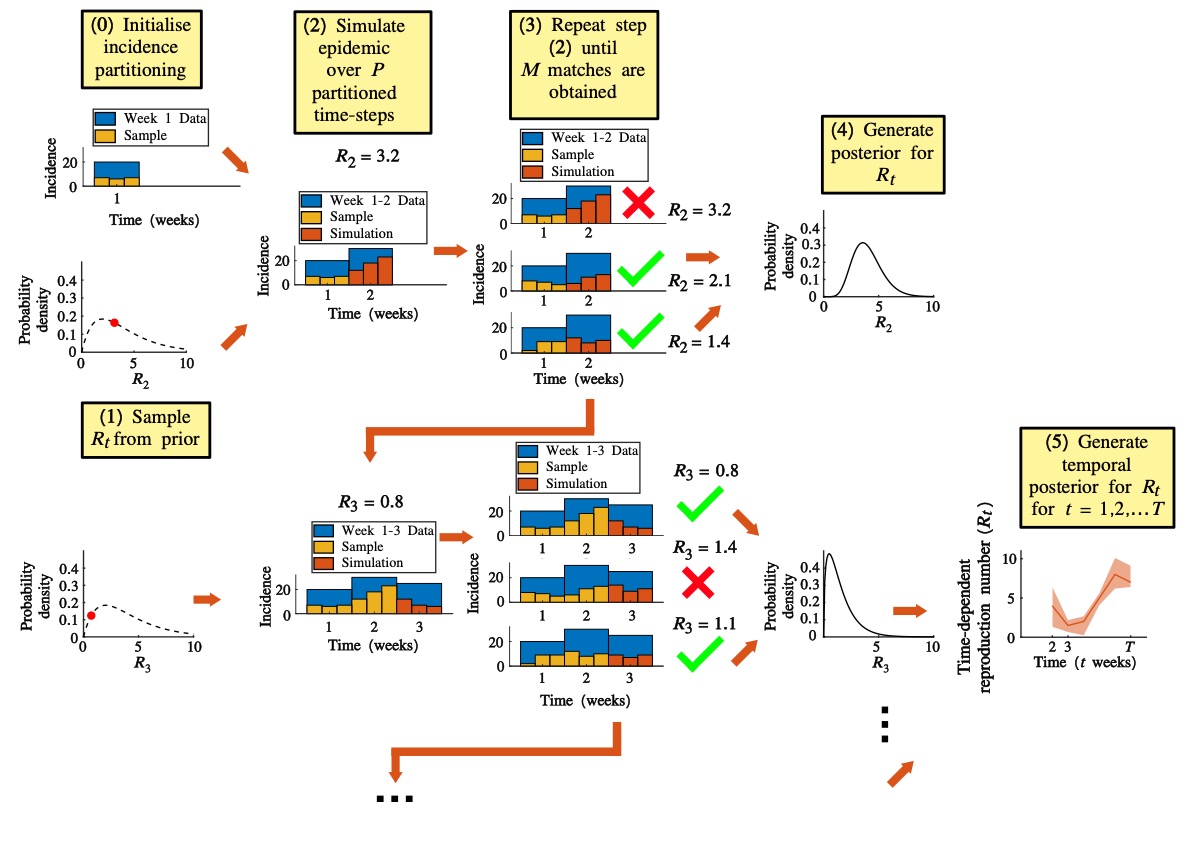
To estimate (for ) from a weekly disease incidence time series dataset, we consider running simulations of the modified renewal equation model in which each week is divided into timesteps (each of timestep weeks). The value therefore corresponds to a daily timestep, however the simulation-based method can be run for any value of (with larger values of leading to the most accurate possible estimates of ).

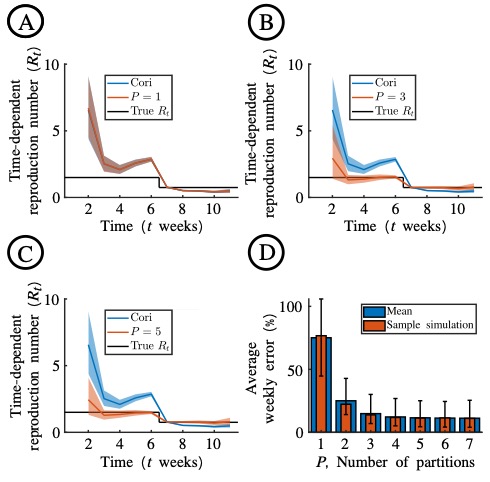
To estimate , we repeatedly simulate the modified renewal equation up until the end of the second week, storing “matching” simulations (those simulations in which the number of cases in the second week in the simulation matches the number of cases in the second week in the time series dataset exactly). In each simulation, we: i) sample the value of from the prior for ; ii) assign each case in the first week of the dataset to one of the timesteps in the first week (chosen uniformly at random). New simulations are generated until simulations that match the number of cases in the second week of the dataset have been obtained. For each matching simulation, we store both the sampled value of and the corresponding numbers of cases in each timestep in that simulation, . The values of from the matching simulations can be combined to construct the posterior distribution for .

We then estimate for each in turn. To do this, we again run simulations of the modified renewal equation model, but starting from the beginning of week (this corresponds to timestep in the modified renewal equation model). Each simulation is run until the end of week (i.e. up to and including timestep ). In each simulation, we: i) sample the value of from the prior; ii) choose past incidence data uniformly at random out of the matching sets stored when estimating . New simulations are generated until simulations that match the number of cases in week of the dataset have been obtained. For each matching simulation, we store both the sampled value of and the corresponding numbers of cases in each timestep in that simulation (including the sampled past incidence data used in that simulation), . The values of from the matching simulations can be combined to construct the posterior distribution for .

In all of our analyses, we required simulations that match the disease incidence time series data in week to have exactly the correct number of cases in that week. For improved computational efficiency, this algorithm could be adapted so that the number of cases in week in matching simulations is within some tolerance level of the corresponding number of cases in the real-world data. However, we did not use that approach here as it is expected to lead to less accurate estimates of , and we found that our computing code ran sufficiently quickly for results to be obtained without this adaptation.

**Supplementary Figures**

  
**Fig S1. Schematic illustrating the steps involved in the simulation-based method for inferring .**

  
**Fig S2. Dependence of estimates using the simulation-based method on the value of used, for the simulated disease incidence time series dataset.** A. Estimates of obtained when the Cori method (blue) and the novel simulation-based approach with (red) are applied to the simulated disease incidence time series dataset (Fig 2A). B. Analogous to panel A, but with in the simulation-based approach. C. Analogous to panel A, but with in the simulation-based approach. D. The average weekly error in estimates obtained using the simulation-based method with different values of *,* compared to the true underlying value of . Red bars are for the simulated dataset shown in Fig 2A of the main text. Blue bars are the mean weekly error in each of 100 simulated datasets that were generated in an identical fashion to the simulated dataset in Fig 2A of the main text. Error bars show the 90% credible interval across the 100 simulations.

**References**

1. 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: 1505–12.

2. Minter A, Retkute R. Approximate Bayesian Computation for infectious disease modelling. Epidemics. 2019;29: 100368.