

## RESEARCH ARTICLE

## Fast estimation of time-varying infectious disease transmission rates

Mikael Jagan<sup>1,2</sup>, Michelle S. deJonge<sup>1</sup>, Olga Krylova<sup>1</sup>, David J. D. Earn<sup>1,2\*</sup>

**1** Department of Mathematics & Statistics, McMaster University, Hamilton, Ontario, Canada, **2** M.G. DeGroot Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada

 These authors contributed equally to this work.

<sup>✉</sup> Current address: Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada

<sup>✉</sup> Current address: Integrated Decision Support, Hamilton Health Sciences, Hamilton, Ontario, Canada

<sup>✉</sup> Current address: Advanced Analytics, Canadian Institute for Health Information, Ottawa, Ontario, Canada

\* [earn@math.mcmaster.ca](mailto:earn@math.mcmaster.ca)



## Abstract

Compartmental epidemic models have been used extensively to study the historical spread of infectious diseases and to inform strategies for future control. A critical parameter of any such model is the transmission rate. Temporal variation in the transmission rate has a profound influence on disease spread. For this reason, estimation of time-varying transmission rates is an important step in identifying mechanisms that underlie patterns in observed disease incidence and mortality. Here, we present and test fast methods for reconstructing transmission rates from time series of reported incidence. Using simulated data, we quantify the sensitivity of these methods to parameters of the data-generating process and to misspecification of input parameters by the user. We show that sensitivity to the user's estimate of the initial number of susceptible individuals—considered to be a major limitation of similar methods—can be eliminated by an efficient, “peak-to-peak” iterative technique, which we propose. The method of transmission rate estimation that we advocate is extremely fast, for even the longest infectious disease time series that exist. It can be used independently or as a fast way to obtain better starting conditions for computationally expensive methods, such as iterated filtering and generalized profiling.

 OPEN ACCESS

**Citation:** Jagan M, deJonge MS, Krylova O, Earn DJD (2020) Fast estimation of time-varying infectious disease transmission rates. *PLoS Comput Biol* 16(9): e1008124. <https://doi.org/10.1371/journal.pcbi.1008124>

**Editor:** Jane M. Heffernan, York University, CANADA

**Received:** September 5, 2019

**Accepted:** July 6, 2020

**Published:** September 21, 2020

**Copyright:** © 2020 Jagan et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are found within the manuscript and Supporting Information files. Specifically, all of the data are simulated and can be generated reproducibly by running the R scripts contained in the Supporting Information files.

**Funding:** MJ was supported by an Undergraduate Student Research Award from the Natural Sciences and Engineering Research Council of Canada (NSERC) and a Student Fellowship from the M. G. DeGroot Institute for Infectious Disease Research. OK was supported by a Postgraduate Scholarship

## Author summary

Many pathogens cause recurrent epidemics. Patterns of recurrence are strongly affected by seasonality of the transmission rate, which can arise from seasonal changes in weather and host population behaviour (e.g., aggregation of children in schools). To understand and predict recurrent epidemic patterns, it is essential to reconstruct the time-varying transmission rate, which is never observed directly. Existing transmission rate estimation methods tend to fall into one of two categories: accurate but too slow to apply to long time series of reported incidence, or fast but inaccurate unless the number of individuals initially susceptible to infection is known precisely. Here, we introduce and compare fast methods inspired by the algorithm that Fine and Clarkson pioneered in the early 1980s.

from NSERC. DJDE was supported by a Discovery Grant from NSERC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

The method that we suggest accurately reconstructs seasonally varying transmission rates, even with crude information about the initial size of the susceptible population.

## 1 Introduction

The transmission rate of an infectious disease is a salient quantity in the study of epidemics. Changes in the transmission rate over time greatly influence the spread of infection [1, 2]. Quantifying how it changes over time can elucidate factors governing disease spread (*e.g.*, weather [3], contact patterns [4]), inform epidemic forecasts, and suggest strategies for epidemic control [5].

In practice, we do not observe transmission directly. Instead, we observe the number of cases of infection (disease incidence) or number of deaths from infection (disease mortality) reported over time, and must reconstruct time-varying transmission rates from these data [6–13]. Utilizing historical mortality records, it is possible to identify patterns in transmission dating far back in time. Most notably, the London Bills of Mortality and the Registrar General's Weekly Returns enable investigation of transmission patterns continuously from the mid-17th century to the present, for a number of infectious diseases including cholera [14] and smallpox [15].

A mechanistic understanding of long infectious disease time series—three centuries of weekly data in the case of smallpox [15]—requires methods of transmission rate estimation that are both accurate and fast, and therefore tractable for long time scales. Simulation-based methods of transmission rate estimation from reported incidence or mortality have been developed using the susceptible-infected-removed (SIR) model for infectious disease dynamics [16]. Markov chain Monte Carlo (MCMC [17, 18]) and sequential Monte Carlo (as in iterated filtering [8, 19, 20]) methods are statistically rigorous, but not tractable for long time scales owing to high computational cost. Generalized profiling [21, 22], which combines trajectory and gradient matching, is faster, but still too slow for convenient exploration of time series spanning hundreds of years. (Several CPU hours were required to apply generalized profiling to just 26 years of weekly data [22].)

In comparison, Finkenstädt and Grenfell's popular “time series SIR” (tSIR) method [7, 23] is extremely fast, using a simple discretization of a continuous-time SIR model to reduce transmission rate estimation to a local regression problem. However, the tSIR method assumes that the duration of infection is equal to the time step, that natural death of susceptible individuals can be ignored, and that cumulative incidence approximates cumulative births. The latter two assumptions are reasonable for pre-vaccination measles, when most susceptible individuals were eventually infected [6]. However, in many contexts (*e.g.*, with pathogens less transmissible than measles), susceptible mortality over time scales of interest and the difference between incidence and births are non-negligible.

In their unpublished PhD and MSc theses, Krylova (Ch. 4 in [24]) and deJonge [25] separately modified a fast discretization method originally proposed by Fine and Clarkson [6]. Krylova's approach has been employed to estimate the amplitude of seasonal variation in measles transmission in 20th century New York City [9]. Unlike the tSIR method and unlike Fine and Clarkson, Krylova's and deJonge's methods do not place constraints on the infectious period or ignore susceptible mortality.

Here, we present a new algorithm based on deJonge's method and compare its performance to the methods of Fine and Clarkson and Krylova. Our main investigative approach is to apply

each method to simulated reported incidence data with known underlying transmission rate, so that error in transmission rate estimates can be quantified exactly.

Our analysis of the methods reveals a shared sensitivity to process and observation error. We mitigate this issue by introducing a smoothing step. The methods are additionally sensitive to error in the user's estimate of the initial number of susceptible individuals, which is rarely known with any precision. We propose a fast, iterative technique for estimating this parameter from time series of incidence, births, and natural mortality, eliminating a long-standing barrier to the use of fast methods of transmission rate reconstruction.

## 2 Methods

In §§2.1 and 2.2 below, we present three fast methods for estimating time-varying transmission rates, based on a mechanistic model of disease spread. In §§2.3–2.7, we outline our systematic analysis of the sensitivity of the methods to parameters of the data-generating process and to error in the user-specified values of input parameters. Finally, in §2.8, we introduce peak-to-peak iteration (PTPI), a technique for estimating the initial number of susceptible individuals. Essential notation is summarized in [Table 1](#).

### 2.1 Model of disease transmission

We assume that the principal mechanisms of disease spread in the focal population are captured by the SIR model [16], formulated with time-varying rates of birth, death, and transmission. Expressing the model as a system of ordinary differential equations, we write

$$\frac{dS}{dt} = v(t)\hat{N}_0 - \beta(t)SI - \mu(t)S, \quad (1a)$$

$$\frac{dI}{dt} = \beta(t)SI - \gamma I - \mu(t)I, \quad (1b)$$

$$\frac{dR}{dt} = \gamma I - \mu(t)R, \quad (1c)$$

where  $S$ ,  $I$ , and  $R$  are the numbers of individuals who are susceptible, infected, and removed, respectively;  $N = S + I + R$  is the population size; and  $\hat{N}_0 = N(0)$  is the population size at an initial time, defined to be 0 years for simplicity. (We reserve the notation  $N_0$  for  $N(t_0)$ , where  $t_0 > 0$  years is the length of a transient; see [Table 1](#).)

The time-varying parameters are

$v(t)$  birth rate, the number of births per unit time relative to  $\hat{N}_0$ ;

$\mu(t)$  natural mortality rate, the number of natural deaths per unit time *per capita* (i.e., relative to  $N$ ); and

$\beta(t)$  transmission rate, the number of infections per unit time per susceptible per infected.

The constant parameter  $\gamma$  is the rate of removal from the infected compartment (due to recovery or death from disease) per infected individual.

In [Eq \(1a\)](#) and [Eq \(1b\)](#), we use mass action incidence  $\beta(t)SI$  rather than standard incidence  $\beta(t)SI/N$ . Mass action incidence is essential for reproducing transitions in epidemic patterns resulting from changes in the birth rate [2, 28]. In [Eq \(1a\)](#), we write the net birth rate as  $v(t)\hat{N}_0$  rather than  $v(t)$ . This formulation is for convenience: the factor of  $\hat{N}_0$  does not affect dynamics, but ensures that  $v(t)$  and  $\mu(t)$  have the same scale.

**Table 1. Notation.** Unless otherwise stated, simulations of reported incidence time series use the reference values listed here. If a symbol is to be interpreted differently in relation to disease incidence and disease mortality data, then the correct definition is indicated by (I) and (M), respectively.

Symbol	Name	Definition	Ref. val.	Unit	Notes
$t_k$	$k$ th observation time	Time of the $k$ th observation in time series data, for $k = 0, \dots, n$ .	$t_0 + k\Delta t$	years	
$t_0$	Transient period	Duration of the transient in system (1) that is ignored in simulations of reported incidence, before observations are recorded.	2000	years	System (1) is numerically integrated between $t = 0$ years and $t = t_0$ , and observed starting at $t = t_0$ . This is done so that simulations reflect dynamics near the attractor of system (1).
$\Delta t$	Observation interval	Time between successive observations in time series data.	1	weeks	Disease mortality is reported weekly in the London Bills of Mortality.
$n$	Time series length	Time between the initial and final observations in time series data, in units $\Delta t$ , given by $(t_n - t_0)/\Delta t$ .	1042	—	If $\Delta t = 1$ week, then $1042\Delta t = \lfloor 20 \times 365/7 \rfloor \Delta t \simeq 20$ years.
$[\cdot]_{\Delta t}$	Nearest $k\Delta t$ rounding	For time lengths $t$ , $[t]_{\Delta t} = \lfloor \frac{t}{\Delta t} \rfloor \Delta t$ , where $[\cdot]$ denotes nearest integer rounding.	—	—	
$\langle \cdot \rangle$	Long-term averaging	For functions $x(t)$ , $\langle x \rangle = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(s) ds$ . For sequences $x_k$ , $\langle x \rangle = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n x_k$ .	—	—	
$(S(t), I(t), R(t))$	State	Number of (susceptible, infected, removed) individuals in the population at time $t$ .	—	—	“Removed” individuals have either recovered from the disease and gained permanent immunity or died from the disease.
$N(t)$	Population size	$S(t) + I(t) + R(t)$ .	—	—	
$B(t)$	Births	Number of births that occur during the time interval $[t - \Delta t, t)$ .	—	—	
$Q(t)$	Cumulative incidence	Number of susceptibles who become infected during the time interval $[t_0, t)$ .	—	—	
$Z(t)$	Incidence	Number of susceptibles who become infected during the time interval $[t - \Delta t, t)$ .	$Q(t) - Q(t - \Delta t)$	—	
$C(t)$	Reported disease (I) incidence or (M) mortality	Number of (I) infections or (M) disease-induced deaths reported during the time interval $[t - \Delta t, t)$ .	Eq (31)	—	C is an abbreviation of “cases”, which are reported as infections or as deaths.
$(S_0, I_0, R_0)$	Initial state ( $t = t_0$ )	$(S(t_0), I(t_0), R(t_0))$ .	$(S^*, I^*, R^*)$	—	$(S^*, I^*, R^*)$ denotes the state of system (1) after numerical integration between $t = 0$ years and $t = t_0$ with seasonally forced transmission rate $\beta(t)$ (Eq (27)), constant vital rates $v_c$ and $\mu_c$ , and initial state ( $t = 0$ years) $(\hat{S}, \hat{I}, \hat{R})$ (Eq (32); see below).
$N_0$	Initial population size ( $t = t_0$ )	$N(t_0)$ .	$S^* + I^* + R^*$	—	
$(\hat{S}, \hat{I}, \hat{R})$	Endemic equilibrium	Endemic equilibrium of system (1) with constant transmission rate ( $\beta \equiv \langle \beta \rangle$ ) and constant vital rates ( $v \equiv \mu \equiv \mu_c$ ).	Eq (32)	—	
$\hat{N}_0$	Initial population size ( $t = 0$ years)	$N(0)$ .	$10^6$	—	
$x_k$	Estimation input/output	Within an estimation algorithm (Boxes 1–3), the supplied or estimated value of $x(t_k)$ ( $x = C, B, \mu, Z, S, I, \beta$ ).	—	varies	
$v(t)$	Birth rate	Number of births per unit time relative to $\hat{N}_0$ , at time $t$ .	$v_c$	year <sup>-1</sup>	In simulations of reported incidence, $v(t)$ is modeled as a constant $v_c$ . In general, estimation of $\beta(t)$ from data does not require the underlying $v(t)$ to be constant.
$v_c$	Birth rate (constant)	See $v(t)$ above.	0.04	year <sup>-1</sup>	
$\mu(t)$	Natural mortality rate	Number of natural deaths per unit time <i>per capita</i> , at time $t$ .	$\mu_c$	year <sup>-1</sup>	In simulations of reported incidence, $\mu(t)$ is modeled as a constant $\mu_c$ . In general, estimation of $\beta(t)$ from data does not require the underlying $\mu(t)$ to be constant.
$\mu_c$	Natural mortality rate (constant)	See $\mu(t)$ above.	0.04	year <sup>-1</sup>	

(Continued)

Table 1. (Continued)

Symbol	Name	Definition	Ref. val.	Unit	Notes
$t_{\text{gen}}$	Mean generation interval	Mean time between onset of infection (in infector) and subsequent transmission of infection (by infector) [26, 27].	13	days	The reference value is the sum of the observed mean latent and infectious periods, which for measles are 8 days and 5 days, respectively [16].
$\gamma$	Removal rate	Number of removals (recoveries or deaths from disease) per unit time per infected.	$1/t_{\text{gen}}$	day <sup>-1</sup>	
$p_{\text{rep}}$	Case reporting probability	(I) Probability that an infection is reported, or (M) the case fatality ratio times the probability that a death from disease is reported.	0.25 or 1	—	If we simulate data with under-reporting, then we use $p_{\text{rep}} = 0.25$ as a reference value. Otherwise, we set $p_{\text{rep}} = 1$ .
$t_{\text{rep}}$	Mean case reporting delay	Mean time between infection and reporting of (I) infection or (M) disease-induced death.	2 or 0	weeks	If we simulate data with reporting delays, then we use $t_{\text{rep}} = 2$ weeks as a reference value. Otherwise, we set $t_{\text{rep}} = 0$ weeks.
$\beta(t)$	Transmission rate	Number of infections per unit time per susceptible per infected, at time $t$ .	Eq (27)	year <sup>-1</sup>	In simulations of reported incidence, $\beta(t)$ is modeled by the seasonal forcing function defined in Eq (27). In general, estimation of $\beta(t)$ from data does not require it vary seasonally or even periodically.
$\langle\beta\rangle$	Mean transmission rate	Continuous-time average of the seasonally forced $\beta(t)$ , equal to $\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \beta(s) ds$ .	$\beta^*$	year <sup>-1</sup>	$\beta^* \approx 5.6 \times 10^{-4} \text{year}^{-1}$ is the value of $\langle\beta\rangle$ that satisfies Eq (2) with $\mathcal{R}_0 = 20$ , $\nu_c = \mu_c = 0.04 \text{year}^{-1}$ , $t_{\text{gen}} = \gamma^{-1} = 13$ days, and $\hat{N}_0 = 10^6$ .
$\mathcal{R}_0$	Basic reproduction number	Number of individuals that a typical infected person is expected to infect in an otherwise completely susceptible population.	Eq (2)	—	For measles in the 20th century, $\mathcal{R}_0 \approx 20$ [16].
$\alpha$	Seasonal amplitude	Amplitude of the seasonally forced $\beta(t)$ relative to $\langle\beta\rangle$ .	0.08	—	For measles, $\alpha \approx 0.08$ [2]. We require $\alpha \in [0, 1]$ to ensure that the seasonal forcing function defined in Eq (27) is non-negative.
$\phi(t; \epsilon)$	Environmental noise (realized)	Phase shift in the seasonally forced $\beta(t)$ , at time $t$ .	Normal(0, $\epsilon^2$ )	—	$\phi$ is a realization of a continuous-time stochastic process defined by a set $\{\Phi(t; \epsilon)\}$ of independent and Normal(0, $\epsilon^2$ )-distributed random variables.
$\epsilon$	Standard deviation of environmental noise	See $\phi(t; \epsilon)$ above.	0.5	—	
$q$	Loess smoothing parameter	Rough number of nearest neighbours weighted in local regression (i.e., when fitting loess curves to time series), determining the degree of smoothing.	—	—	See §2.2.6 for an exact definition.

<https://doi.org/10.1371/journal.pcbi.1008124.t001>

The SIR model (1) assumes that the focal population is homogeneously mixed and subject to the mass action principle, which states that incidence is proportional to the product of the densities of susceptibles and infecteds [16]. The model further assumes that the latent period (time from infection to onset of infectiousness) can be ignored and that the infectious period (time from onset of infectiousness to recovery or death from disease) is exponentially distributed [29]. The distributions of the latent and infectious periods affect disease dynamics [28, 30, 31], but Krylova and Earn [28] showed that the effect on long-term dynamical structure is typically small when the mean generation interval is fixed (see Fig 11 in [28]). For this reason, we assign the mean generation interval implied by the SIR model (1) ( $t_{\text{gen}} = \gamma^{-1}$ ) the value of the sum of the observed mean latent and infectious periods. This sum is the true mean generation interval if the latent and infectious periods are both exponentially distributed, and is a good estimate of the true mean generation interval more generally [28].

Transmissibility of infection is typically measured by the basic reproduction number  $\mathcal{R}_0$ , defined as the number of individuals that a typical infected person is expected to infect in an otherwise completely susceptible population [16]. If the birth and death rates are constant ( $\nu \equiv \nu_c$  and  $\mu \equiv \mu_c$ ), and if the transmission rate has a well-defined average  $\langle\beta\rangle$  [32], then the

basic reproduction number for the SIR model (1) can be written as [28]

$$\mathcal{R}_0 = \frac{v_c \hat{N}_0}{\mu_c} \cdot \frac{\langle \beta \rangle}{\gamma + \mu_c}. \quad (2)$$

## 2.2 Estimating $\beta(t)$ from time series data

Here, we examine three fast methods for estimating time-varying transmission rates  $\beta(t)$ . The methods take as input (i) time series of reported disease incidence or disease mortality, (ii) time series of births and natural mortality, and (iii) values for input parameters, such as the mean generation interval  $t_{\text{gen}}$ . By assumption, the time series data are available at discrete, equally spaced time points

$$t_k = t_0 + k\Delta t, \quad k = 0, \dots, n, \quad (3)$$

where  $\Delta t$  is the observation interval. The methods return as output a time series estimate of  $\beta(t)$ , denoted by  $\{(t_k, \beta_k)\}_{k=0}^n$  or simply  $\beta_k$ , which can be averaged (§2.2.5) or smoothed (§2.2.6) to distill temporal patterns of interest.

Missing data must be imputed: the three methods are recursive, so they break down as soon as they encounter a missing value. Imputation can be accomplished most simply via linear interpolation between available data. More sophisticated techniques accounting for uncertainty in missing values are described in [33].

**2.2.1 FC method.** We review the method first described by Fine and Clarkson [6], referred to here as the “FC method”. Let  $S(t)$  and  $I(t)$  be the number of susceptibles and infecteds in the population at time  $t$ .  $S$  decreases when susceptibles become infected or die and increases when susceptibles are born. Let  $Z(t)$  and  $B(t)$  be the number of infections and births, respectively, that occur during the time interval  $[t - \Delta t, t)$ . Assuming that natural mortality was negligible, Fine and Clarkson reconstructed  $S$  from  $Z$  and  $B$  with the recursion

$$S(t + \Delta t) \approx S(t) + B(t + \Delta t) - Z(t + \Delta t). \quad (4)$$

Fine and Clarkson further assumed that the observation interval  $\Delta t$  was equal to the mean generation interval  $t_{\text{gen}}$ , so that prevalence could be approximated by incidence. That is,

$$I(t) \approx Z(t) \quad (5)$$

for all  $t$ . They derived an expression for  $Z(t + \Delta t)$  via the mass action principle

$$Z(t + \Delta t) \approx \beta(t)S(t)I(t)\Delta t. \quad (6)$$

Rearranging Eq (6), they obtained an estimate of  $\beta(t)$ , given by

$$\beta(t) \approx \frac{Z(t + \Delta t)}{S(t)I(t)\Delta t}. \quad (7)$$

Fine and Clarkson applied Eqs (4), (5), and (7) to estimate  $S(t_k)$ ,  $I(t_k)$ , and  $\beta(t_k)$  (for  $k = 0, \dots, n$ ), after specifying (i) the initial number of susceptibles  $S_0 = S(t_0)$ , and (ii) values of  $Z(t_k)$  and  $B(t_k)$  from incidence and birth data, respectively.

A limitation of the FC method is the constraint requiring  $\Delta t = t_{\text{gen}}$ . For some diseases, this is a minor issue, because incidence and birth data can be aggregated so that the time between successive aggregates is approximately equal to  $t_{\text{gen}}$ . For example, the mean generation interval of measles is approximately two weeks, so Fine and Clarkson [6] aggregated pairs of weekly

observations. A second, more serious limitation is the assumption, implicit in Eqs (4) and (5), that natural mortality is negligible. We discuss this issue in §3.1.

**2.2.2 S method.** Krylova (Ch. 4 in [24]) generalized the FC method in order to eliminate the constraint requiring  $\Delta t = t_{\text{gen}}$  and account for natural mortality. Her approach is based on the SEIR model, which distinguishes “exposed” individuals in the latent stage of infection from infectious individuals. Here, we adapt Krylova’s approach to the SIR model (1) and refer to our approach as the “S method.”

We define  $S, I, Z,$  and  $B$  as in the FC method. Let  $\mu(t)$  be the *per capita* natural mortality rate at time  $t$ , and let  $Q(t)$  be the total number of infections occurring between the initial observation time  $t_0$  and current time  $t$  (*i.e.*, cumulative incidence). The observation interval  $\Delta t$  is no longer constrained to be equal to the mean generation interval  $t_{\text{gen}}$ .

We reconstruct  $S$  recursively by discretizing Eq (1a):

$$S(t + \Delta t) \approx S(t) + B(t + \Delta t) - Z(t + \Delta t) - \mu(t)S(t)\Delta t. \tag{8}$$

Eq (8) is the result of applying the forward Euler method for numerical integration to Eq (1a), and replacing the incidence and birth terms with  $Z(t + \Delta t)$  and  $B(t + \Delta t)$ , respectively. Eq (8) is identical to Eq (4) of the FC method, except that it includes a natural mortality term.

In order to estimate  $\beta(t)$ , we note that, by definition,  $dQ/dt$  is the rate at which individuals enter the infected compartment. From Eq (1b), this is

$$\frac{dQ}{dt} = \beta(t)S(t)I(t). \tag{9}$$

If the mean generation interval  $t_{\text{gen}}$  is short enough that  $I$  and  $\mu$  are roughly constant between times  $t$  and  $t + t_{\text{gen}}$ , then  $dI/dt \approx 0$  in that interval, and using Eq (1b) we can write

$$\beta(t)S(t)I(t) \approx (\gamma + \mu(t))I(t) \approx (\gamma + \mu(t + t_{\text{gen}}))I(t + t_{\text{gen}}). \tag{10}$$

In this case,  $dQ/dt$  is also (approximately) the rate at which individuals leave the infected compartment,  $t_{\text{gen}}$  time after infection:

$$\frac{dQ}{dt} \approx (\gamma + \mu(t + t_{\text{gen}}))I(t + t_{\text{gen}}). \tag{11}$$

Note that Eq (11) is also valid if the generation interval is narrowly distributed around its mean  $t_{\text{gen}}$  (even if  $t_{\text{gen}}$  is long).

Discretizing Eqs (9) and (11) using forward Euler, we obtain two approximations of  $Z(t + \Delta t)$ :

$$Z(t + \Delta t) = Q(t + \Delta t) - Q(t) \approx \begin{cases} \beta(t)S(t)I(t)\Delta t & \text{from Eq (9),} \\ (\gamma + \mu(t + t_{\text{gen}}))I(t + t_{\text{gen}})\Delta t & \text{from Eq (11).} \end{cases} \tag{12}$$

Rearranging Eq (12) yields an estimate of  $\beta(t)$ , given by

$$\beta(t) \approx \frac{Z(t + \Delta t)}{S(t)I(t)\Delta t}, \tag{13}$$

and an estimate of  $I(t)$ , given by

$$I(t) \approx \frac{Z(t + \Delta t - t_{\text{gen}})}{(\gamma + \mu(t))\Delta t}. \tag{14}$$

Since data are available only at the observation times  $t_k$  (Eq (3)), the value of  $Z(t + \Delta t - t_{\text{gen}})$  in



Eq (14) will be observed only if  $t_{\text{gen}}$  is an integer multiple of  $\Delta t$ . In general,  $t_{\text{gen}}$  is not divisible by  $\Delta t$ . Therefore, in practice, we replace  $t_{\text{gen}}$  in  $Z(t + \Delta t - t_{\text{gen}})$  with the nearest integer multiple of  $\Delta t$ , denoted here by  $[t_{\text{gen}}]_{\Delta t}$ :

$$I(t) \approx \frac{Z(t + \Delta t - t_{\text{gen}})}{(\gamma + \mu(t))\Delta t}. \tag{15}$$

Thus, the S method is defined by Eq (13), coupled with Eqs (8) and (15) for the reconstruction of  $S$  and  $I$ . The S method requires users to specify (i) input parameters  $S_0 = S(t_0)$  and  $t_{\text{gen}} = \gamma^{-1}$ , and (ii) values of  $Z(t_k)$ ,  $B(t_k)$ , and  $\mu(t_k)$  from incidence, birth, and natural mortality data, respectively.

The FC method is a special case of the S method, obtained by setting  $\Delta t = t_{\text{gen}}$  and  $\mu(t) \equiv 0$ .

**2.2.3 SI method.** DeJonge [25] improved Krylova’s method (Ch. 4 in [24]) by reconstructing  $I$  directly from Eq (1b) instead of relying on the approximation in Eq (11). Here, we improve deJonge’s discretization, which employs the forward Euler method, by instead combining forward and backward Euler. One way to do this is to use the trapezoidal method: whereas forward and backward Euler take  $f'(t)\Delta t$  and  $f'(t + \Delta t)\Delta t$ , respectively, to approximate integrals  $\int_t^{t+\Delta t} f'(\tau) d\tau$ , the trapezoidal method takes the average  $\frac{1}{2}[f'(t) + f'(t + \Delta t)]\Delta t$ , which is less prone to error. Our discretization, which we call the “SI method”, is consistently more accurate than deJonge’s and others (see §S9 of S1 Text for a comparison of nine possible algorithms). Numerically integrating Eq (1a) and Eq (1b) using the trapezoidal method, and replacing the incidence and birth terms with  $Z(t + \Delta t)$  and  $B(t + \Delta t)$ , respectively, we obtain

$$S(t + \Delta t) \approx \frac{[1 - \frac{1}{2}\mu(t)\Delta t]S(t) + B(t + \Delta t) - Z(t + \Delta t)}{1 + \frac{1}{2}\mu(t + \Delta t)\Delta t} \tag{16}$$

and

$$I(t + \Delta t) \approx \frac{[1 - \frac{1}{2}(\gamma + \mu(t))\Delta t]I(t) + Z(t + \Delta t)}{1 + \frac{1}{2}(\gamma + \mu(t + \Delta t))\Delta t}. \tag{17}$$

Eq (17) eliminates an important problem with Eq (15) of the S method, which estimates  $I(t) \approx 0$  if  $Z(t + \Delta t - [t_{\text{gen}}]_{\Delta t}) = 0$ , leading to division by zero in Eq (13).

Discretizing Eq (9) using forward and backward Euler, we obtain two approximations of  $Z(t + \Delta t)$ :

$$Z(t + \Delta t) = Q(t + \Delta t) - Q(t) \approx \begin{cases} \beta(t)S(t)I(t)\Delta t & \text{from forward Euler,} \\ \beta(t + \Delta t)S(t + \Delta t)I(t + \Delta t)\Delta t & \text{from backward Euler.} \end{cases} \tag{18}$$

Rearranging Eq (18) yields two estimates of  $\beta(t)$ ,

$$\beta(t) \approx \begin{cases} \frac{Z(t+\Delta t)}{S(t)I(t)\Delta t} & \text{from forward Euler,} \\ \frac{Z(t)}{S(t)I(t)\Delta t} & \text{from backward Euler,} \end{cases} \tag{19}$$

whose average supplies a more accurate estimate (see §S9 of S1 Text), given by

$$\beta(t) \approx \frac{Z(t) + Z(t + \Delta t)}{2S(t)I(t)\Delta t}. \tag{20}$$

Thus, the SI method is defined by Eq (20), coupled with Eqs (16) and (17) for the reconstruction of  $S$  and  $I$ . Compared to the S method, the SI method, in principle, requires one



additional input parameter, namely the initial number of infecteds  $I_0 = I(t_0)$ . In §3.6, we show that, in practice, this additional information is not necessary.

**2.2.4 Estimating true incidence from reported incidence.** Let  $C(t)$  be the number of infections reported during the time interval  $[t - \Delta t, t)$ . We estimate true incidence  $Z$  from reported incidence  $C$  via

$$Z(t) \approx \frac{1}{p_{\text{rep}}} C(t + [t_{\text{rep}}]_{\Delta t}), \tag{21}$$

where  $p_{\text{rep}}$  is the probability that an infection is reported and  $[t_{\text{rep}}]_{\Delta t}$  is the mean time between infection and reporting, rounded to the nearest integer multiple of the observation interval  $\Delta t$ .

Eq (21) has the limitation that multiplying by  $p_{\text{rep}}^{-1}$  does not correct for under-reporting if, by chance,  $C(t + [t_{\text{rep}}]_{\Delta t}) = 0$ . In this situation, not only is the result  $Z(t) \approx 0$  incorrect, but we divide by zero in the FC and S methods when we substitute Eqs (5) and (15) in Eqs (7) and (13), respectively. If  $C(t + [t_{\text{rep}}]_{\Delta t}) = C(t + [t_{\text{rep}}]_{\Delta t} + \Delta t) = 0$ , then the SI method also suffers: Eq (20) gives  $\beta(t) \approx 0$ . To circumvent these issues, we replace zeros in reported incidence time series by linearly interpolating between nonzero values prior to estimating true incidence using Eq (21). We do not replace leading and trailing zeros.

If what we observe is deaths from disease, rather than infections, then we have the complication that only a fraction of infections end in death. In this situation, we can still use Eq (21) to estimate  $Z$ , provided we interpret (i)  $C$  as reported disease mortality, (ii)  $p_{\text{rep}}$  as the case fatality ratio times the probability that a death from disease is reported, and (iii)  $t_{\text{rep}}$  as the mean time between infection and reporting of disease-induced death.

A more sophisticated method of inferring true incidence from reported data is described in [34].

**2.2.5 Averaging raw estimates of  $\beta(t)$ .** Given fixed time series data and input parameters, the FC, S, and SI methods return estimates of  $\beta(t)$  that are entirely determined (not random). In the absence of additional data observed from the same population, it is difficult to assign confidence to the output.

However, if an estimate  $\tilde{\beta}(t)$  is approximately periodic (with apparent period  $T$ ) and contains  $m$  complete cycles, and if we assume  $\beta(t)$  is truly periodic, then we can view  $\tilde{\beta}(t)$  as containing a sample of  $m$  estimates of the true cycle, with some variance, and use its mean as an estimator instead of any one of the  $m$  cycles. For such an estimate  $\tilde{\beta}(t)$  defined on the interval  $[t_0, t_0 + mT)$ , the mean and variance are given by

$$\bar{x}(t) = \frac{1}{m} \sum_{i=0}^{m-1} \tilde{\beta}(t + iT), \quad t \in [t_0, t_0 + T), \tag{22a}$$

$$s^2(t) = \frac{1}{m-1} \sum_{i=0}^{m-1} [\tilde{\beta}(t + iT) - \bar{x}(t)]^2, \quad t \in [t_0, t_0 + T). \tag{22b}$$

In §3.3, we apply the S and SI methods to simulated data to estimate an underlying, seasonally forced  $\beta(t)$  (Eq (27)), which has a period of 1 year. We linearly interpolate the raw time series estimate  $\beta_k$  and compute the average 1-year cycle in the interpolant  $\beta_{\text{int}}(t)$  using Eq (22a).

Comparing this average to the true 1-year cycle, we are able to assess bias in the two methods.

Note that  $\bar{x}(t)$  and  $s^2(t)$  can be used to obtain a formal, likelihood-based measure of confidence in estimates  $\tilde{\beta}(t)$  (see §2.3.4 in [35]).

**2.2.6 Smoothing raw estimates of  $\beta(t)$ .** Process and observation error introduce random fluctuations in reported incidence on top of longer-term (e.g., seasonal) variation. In §3.2, we show that noise in reported incidence is spuriously amplified in  $\beta_k$ , the raw time series estimate of  $\beta(t)$ .

To distill temporal patterns of interest from the noise, we fit a smooth loess (short for local regression; see Ch. 8.1 in [36]) curve  $\beta_{\text{loess}}(t; q)$  to the points  $\{(t_k, \beta_k)\}_{k=0}^n$  and use  $\beta_{\text{loess}}(t; q)$  as our final estimate of  $\beta(t)$ . Here,  $q \in \{5, \dots, n + 1\}$  is an integer-valued parameter controlling the degree of smoothing. At times  $t \in [t_0, t_n]$ , the fitted value  $\beta_{\text{loess}}(t; q)$  is obtained as follows:

1. Order the distances  $d_k = |t_k - t|$  of the time points  $t_k$  (Eq (3)) from  $t$ , letting  $d_{k_i}$  denote the  $i$ th smallest distance (for  $i = 1, \dots, n + 1$ ).
2. Fit a quadratic polynomial  $p_2(t)$  to the points  $\{(t_k, \beta_k)\}_{k=0}^n$ . This is done by weighted least squares using tricube weights

$$w_k = \begin{cases} \left(1 - \left(\frac{d_k}{d_{k_q}}\right)^3\right)^3 & \text{if } 0 \leq d_k < d_{k_q}, \\ 0 & \text{if } d_k \geq d_{k_q}. \end{cases} \tag{23}$$

Hence only time points  $t_k$  nearer to  $t$  than the  $q$ th nearest time point are weighted in the fit.

3. Define  $\beta_{\text{loess}}(t; q) = p_2(t)$ .

Typically, smoother fits are obtained with greater  $q$  [36, 37].

The optimal value of  $q$  for a given time series  $\beta_k$ , denoted by  $q_{\text{opt}}$ , is that which minimizes error in  $\beta_{\text{loess}}(t; q)$  relative to  $\beta(t)$ . In §3.4, we estimate  $\beta(t)$  from simulated data, smooth  $\beta_k$  using each value of  $q$  on a grid, and use our knowledge of  $\beta(t)$  to determine  $q_{\text{opt}}$ . We show that it is possible for smoothing to eliminate much of the error in  $\beta_k$  attributable to process and observation error. Thus, in §2.2.7, we explicitly define the FC, S, and SI methods with loess smoothing as a final step.

In practice,  $\beta(t)$  is not known, so we cannot determine  $q_{\text{opt}}$ . In this case,  $q_{\text{opt}}$  can be estimated using statistical methods, such as time series cross-validation [38]. However, reasonable results can be obtained much more quickly by inspecting  $\beta_{\text{loess}}(t; q)$  directly and increasing  $q$  from 4 until a desirable degree of smoothing is achieved (e.g., until noise on the time scale of weeks is visibly reduced, and patterns on the time scale of months are easier to discern).

**2.2.7 Summary.** In Boxes 1–3 below, we summarize the three methods derived in §§2.2.1–2.2.6 for estimating time-varying transmission rates  $\beta(t)$  from time series data with observation times  $t_k$  (Eq (3)). We use the notation  $x_k$  to refer to the value supplied or computed for  $x(t_k)$  within the estimation algorithms ( $x = C, B, \mu, Z, S, I, \beta$ ).

In Box 4, we provide instructions for input specification based on our analysis of the methods.

**Box 1. FC method (Fine & Clarkson 1982 [6])**

$$Z_k \leftarrow \frac{1}{p_{\text{rep}}} C_{k+r}, \quad \text{where } r = \frac{\lceil t_{\text{rep}} \rceil \Delta t}{\Delta t}, \tag{24a}$$

$$S_k \leftarrow S_{k-1} + B_k - Z_k, \tag{24b}$$

$$I_k \leftarrow Z_k, \tag{24c}$$

$$\beta_k \leftarrow \frac{Z_{k+1}}{S_k I_k \Delta t}, \tag{24d}$$

where  $\Delta t$  is assumed to be roughly equal to  $t_{\text{gen}}$ , and natural mortality is assumed to be negligible. Users must specify:

- a time series  $\{(t_k, C_k)\}_{k=0}^n$  of reported incidence or reported disease mortality, with zeros replaced via linear interpolation between nonzero values;
- a time series  $\{(t_k, B_k)\}_{k=0}^n$  of births;
- input parameters  $S_0, t_{\text{gen}}, p_{\text{rep}}$ , and  $t_{\text{rep}}$ .

**Box 2. S method (adapted from Krylova 2011 [24])**

$$Z_k \leftarrow \frac{1}{p_{\text{rep}}} C_{k+r}, \quad \text{where } r = \frac{\lceil t_{\text{rep}} \rceil \Delta t}{\Delta t}, \tag{25a}$$

$$S_k \leftarrow S_{k-1} + B_k - Z_k - \mu_{k-1} S_{k-1} \Delta t, \tag{25b}$$

$$I_k \leftarrow \frac{Z_{k+1-g}}{(\gamma + \mu_k) \Delta t}, \quad \text{where } g = \frac{\lceil t_{\text{gen}} \rceil \Delta t}{\Delta t}, \tag{25c}$$

$$\beta_k \leftarrow \frac{Z_{k+1}}{S_k I_k \Delta t}, \tag{25d}$$

$$\beta_{\text{loess}}(t; q) \leftarrow \text{loess curve fit to time series } \{(t_k, \beta_k)\}_{k=0}^n. \tag{25e}$$

Users must specify:

- a time series  $\{(t_k, C_k)\}_{k=0}^n$  of reported incidence or reported disease mortality, with zeros replaced via linear interpolation between nonzero values;
- a time series  $\{(t_k, B_k)\}_{k=0}^n$  of births;
- a time series  $\{(t_k, \mu_k)\}_{k=0}^n$  of the *per capita* natural mortality rate;
- input parameters  $S_0, t_{\text{gen}} = \gamma^{-1}, p_{\text{rep}}$ , and  $t_{\text{rep}}$ ;
- loess smoothing parameter  $q$ .

## Box 3. SI method (adapted from deJonge 2014 [25])

$$Z_k \leftarrow \frac{1}{p_{\text{rep}}} C_{k+r}, \quad \text{where } r = \frac{\lceil t_{\text{rep}} \rceil \Delta t}{\Delta t}, \quad (26a)$$

$$S_k \leftarrow \frac{[1 - \frac{1}{2}\mu_{k-1}\Delta t]S_{k-1} + B_k - Z_k}{1 + \frac{1}{2}\mu_k\Delta t}, \quad (26b)$$

$$I_k \leftarrow \frac{[1 - \frac{1}{2}(\gamma + \mu_{k-1})\Delta t]I_{k-1} + Z_k}{1 + \frac{1}{2}(\gamma + \mu_k)\Delta t}, \quad (26c)$$

$$\beta_k \leftarrow \frac{Z_k + Z_{k+1}}{2S_k I_k \Delta t}, \quad (26d)$$

$$\beta_{\text{loess}}(t; q) \leftarrow \text{loess curve fit to time series } \{(t_k, \beta_k)\}_{k=0}^n. \quad (26e)$$

Users must specify:

- a time series  $\{(t_k, C_k)\}_{k=0}^n$  of reported incidence or reported disease mortality, with zeros replaced via linear interpolation between nonzero values;
- a time series  $\{(t_k, B_k)\}_{k=0}^n$  of births;
- a time series  $\{(t_k, \mu_k)\}_{k=0}^n$  of the *per capita* natural mortality rate;
- input parameters  $S_0, I_0, t_{\text{gen}} = \gamma^{-1}, p_{\text{rep}}$ , and  $t_{\text{rep}}$ ;
- loess smoothing parameter  $q$ .

## Box 4. Instructions for input specification

- $\beta_k$  is sensitive to mis-specification of  $S_0$ , but not  $I_0$  (cf. §3.6.1). If the user's estimate of  $S_0$  is uncertain, and if the incidence time series  $Z_k$  is roughly periodic, then a more accurate estimate of  $S_0$  may be obtained via peak-to-peak iteration (PTPI; cf. §3.7).
- If  $S_k$  is negative for any  $k$ , then it is likely that the case reporting probability  $p_{\text{rep}}$  was underestimated or that births were systematically under-reported by  $B_k$ . This can be resolved by correcting the estimate of  $p_{\text{rep}}$  or correcting  $B_k$ , then restarting the algorithm. Users should apply close to the minimal correction necessary to prevent negative  $S_k$ .
- $q$  must be tuned to the  $\beta_k$  time series. An estimate of  $q_{\text{opt}}$  can be obtained using statistical methods, such as time series cross-validation [38]. However,  $q$  can be tuned quickly through visual inspection of  $\beta_{\text{loess}}(t; q)$ : one can increase  $q$  from 5 until a desirable degree of smoothing is achieved (e.g., until noise on the time scale of weeks is visibly reduced, and patterns on the time scale of months are easier to discern).

### 2.3 Simulating reported incidence data

In order to compare the performance of the FC, S, and SI methods in estimating  $\beta(t)$ , we apply the methods to simulated reported incidence data with known underlying  $\beta(t)$ . Here, we outline our methods for simulating these data using the SIR model (1).

**2.3.1 Seasonal forcing of  $\beta(t)$  with environmental stochasticity.** We reproduce seasonal fluctuation in the transmission rate by modeling  $\beta(t)$  in Eq (1) as a sinusoidal forcing function with period equal to one year:

$$\beta(t) = \langle \beta \rangle \left( 1 + \alpha \cos \left( \frac{2\pi t}{1 \text{ year}} \right) \right). \quad (27)$$

Here,  $\alpha \in [0, 1]$  is the amplitude of seasonal forcing relative to the mean  $\langle \beta \rangle$ . We introduce stochastic fluctuation by adding a randomly generated phase shift:

$$\beta_\phi(t) = \langle \beta \rangle \left( 1 + \alpha \cos \left( \frac{2\pi t}{1 \text{ year}} + \phi(t; \epsilon) \right) \right). \quad (28)$$

$\phi$  is a realization of a continuous-time stochastic process consisting of independent, Normal(0,  $\epsilon^2$ )-distributed random variables. It models **environmental stochasticity** leading to random noise in the transmission rate. Modeling environmental stochasticity with a random phase shift rather than additive noise conveniently avoids negative  $\beta_\phi(t)$ :  $\beta_\phi(t)$  oscillates between  $\langle \beta \rangle(1 - \alpha)$  and  $\langle \beta \rangle(1 + \alpha)$  regardless of the distribution of the noise. In practice, we take the values of  $\phi$  at times  $t_k$  (Eq (3)) and linearly interpolate to obtain values in between. This helps to make simulations of Eqs (1) and (9) with adaptive time steps (cf. §2.3.2) reproducible.

**2.3.2 Generating incidence time series with demographic stochasticity.** We supplement Eq (1) with Eq (9), so that trajectories of the resulting system record changes in cumulative incidence  $Q$ . In this system, we employ the noisy transmission rate  $\beta_\phi(t)$  (Eq (28)) and constant vital rates  $\nu_c$  and  $\mu_c$ . We then either (i) numerically integrate the differential equations to approximate their solution, or (ii) treat the system more realistically as an event-driven, continuous-time Markov process (with event rates specified by terms in the differential equations) and use the adaptive tau-leaping algorithm for stochastic simulation [39, 40]. The latter approach accounts for **demographic stochasticity** in disease dynamics. We prevent disease fadeout in simulations with demographic stochasticity by setting the rates of infected recovery and death to zero whenever  $I = 1$ .

In both methods of simulation, we record the state of the system at times  $t_k$  (Eq (3)), choosing initial state

$$\begin{pmatrix} S(t_0) \\ I(t_0) \\ R(t_0) \\ Q(t_0) \end{pmatrix} = \begin{pmatrix} S_0 \\ I_0 \\ R_0 \\ 0 \end{pmatrix}, \quad (29)$$

where  $S_0 + I_0 + R_0 = N_0 = N(t_0)$ . Finally, we derive incidence  $Z$  from  $Q$  via first differences:

$$Z(t) = Q(t) - Q(t - \Delta t). \quad (30)$$

**2.3.3 Introducing observation error.** Observation error due to under-reporting ( $p_{\text{rep}} < 1$ ) and reporting delays ( $t_{\text{rep}} > 0$  weeks) creates discrepancies between true incidence  $Z$  and reported incidence  $C$ . We introduce random observation error to simulated incidence time

series with delayed binomial sampling:

$$C(t + [t_{\text{rep}}]_{\Delta t}) \sim \text{Binomial}(Z(t), p_{\text{rep}}). \tag{31}$$

For simulations without observation error, we set  $p_{\text{rep}} < 1$  and  $t_{\text{rep}} > 0$  weeks.

**2.3.4 Parametrization.** The simulation method outlined in §§2.3.1–2.3.3 is parametrized by

disease parameters	$\langle \beta \rangle, \alpha, \epsilon, t_{\text{gen}};$
population parameters	$\hat{N}_0 = N(0), S_0 = S(t_0), I_0 = I(t_0), \nu_c, \mu_c;$ and
reporting parameters	$p_{\text{rep}}, t_{\text{rep}}, t_0, \Delta t, n.$

For most simulations, we assign parameters the reference values listed in Table 1. We consider different values when we investigate the sensitivity of  $\beta(t)$  estimates to data-generating parameters (cf. §2.6.1).

We bypass transient dynamics by choosing  $t_0 = 2000$  years and numerically integrating system (1) between 0 years and  $t_0$  in order to obtain a point  $(S^*, I^*, R^*)$  very near the attractor. For this step, we exclude environmental noise, defining  $\beta(t)$  as in Eq (27), and take the initial state to be the endemic equilibrium of the unforced system (system (1) with  $\beta \equiv \langle \beta \rangle$  and  $\nu \equiv \mu \equiv \mu_c$ ):

$$\begin{pmatrix} S(0) \\ I(0) \\ R(0) \end{pmatrix} = \begin{pmatrix} \hat{S} \\ \hat{I} \\ \hat{R} \end{pmatrix} = \begin{pmatrix} \frac{\hat{N}_0}{\mathcal{R}_0} \\ \hat{N}_0 \left(1 - \frac{1}{\mathcal{R}_0}\right) \left(\frac{\mu_c}{\gamma + \mu_c}\right) \\ \hat{N}_0 \left(1 - \frac{1}{\mathcal{R}_0}\right) \left(\frac{\gamma}{\gamma + \mu_c}\right) \end{pmatrix}. \tag{32}$$

### 2.4 Creating mock birth and natural mortality time series

In addition to reported incidence data, the FC, S, and SI methods require time series of births and the *per capita* natural mortality rate. For simplicity, we create mock time series by (i) choosing constant vital rates  $\nu'_c$  and  $\mu'_c$ , then (ii) setting  $B_k = \nu'_c \hat{N}_0 \Delta t$  and  $\mu_k = \mu'_c$  for all  $k$ . Note that  $\nu'_c \hat{N}_0 \Delta t$  is the result of integrating the net birth rate in the SIR model (1), given by  $\nu(t) \hat{N}_0$ , between successive observation times using  $\nu \equiv \nu'_c$ .

We specify  $\nu'_c = \nu_c$  and  $\mu'_c = \mu_c$ , where  $\nu_c$  and  $\mu_c$  are the data-generating vital rates (cf. §2.3.4), except when we investigate the sensitivity of  $\beta(t)$  estimates to incorrect vital data (cf. §2.6.2). For example, to model under-reporting of births, we simply set  $\nu'_c < \nu_c$ .

### 2.5 Measuring $\beta(t)$ estimation error

When we simulate reported incidence data, the underlying transmission rate  $\beta(t)$  is defined beforehand via Eq (27) and known for all  $t$ . We use this knowledge to quantify the error in estimates of  $\beta(t)$  obtained from the data. Specifically, given an estimate  $\tilde{\beta}(t)$  defined at time points  $t_k$  (Eq (3)), we compute the relative root mean square error (RRMSE), defined as

$$\text{RRMSE}(\beta, \tilde{\beta}) := \sqrt{\frac{1}{n+1} \sum_{k=0}^n \left( \frac{\beta(t_k) - \tilde{\beta}(t_k)}{\tilde{\beta}} \right)^2}, \tag{33}$$

where

$$\bar{\beta} := \frac{1}{n+1} \sum_{k=0}^n \beta(t_k). \quad (34)$$

Note that by “underlying transmission rate” we mean the transmission rate *excluding* environmental noise. Although we simulate data using the noisy  $\beta_\phi(t)$ , defined in Eq (28), our aim is to reconstruct the noiseless  $\beta(t)$ , defined in Eq (27).

## 2.6 Sensitivity analysis

Error in  $\beta(t)$  estimation from reported incidence data depends on how the data were generated. The number of cases reported over time is influenced by features of the disease (e.g., the natural history of infection), population (e.g., contact patterns), and case reporting (e.g., the frequency and accuracy of reports). In our simulations of reported incidence, there are 14 **data-generating parameters** (cf. §2.3.4), whose values are summarized in the vector

$$\theta = (\langle \beta \rangle, \alpha, \epsilon, \hat{N}_0, S_0, I_0, v_c, \mu_c, t_{\text{gen}}, p_{\text{rep}}, t_{\text{rep}}, t_0, \Delta t, n). \quad (35)$$

Estimation error also depends on how accurately certain data-generating parameters are specified by users of the FC, S, and SI methods. The initial observation time  $t_0$ , observation interval  $\Delta t$ , and time series length  $n$  are always known exactly. Other parameters ( $\langle \beta \rangle, \alpha, \epsilon, \hat{N}_0, v_c$ , and  $\mu_c$ ) influence our simulations of reported incidence, but in practice are not parameters of the FC, S, and SI methods. In practice, users are required to specify only  $S_0, t_{\text{gen}}, p_{\text{rep}}, t_{\text{rep}}$ , and (with the SI method)  $I_0$ . However, when we test the methods here, we do specify vital rates  $v_c$  and  $\mu_c$  in order to create mock (constant) birth and natural mortality time series (cf. §2.4). The specified values of these 7 **input parameters** are summarized in the vector

$$\theta' = (S'_0, I'_0, v'_c, \mu'_c, t'_{\text{gen}}, p'_{\text{rep}}, t'_{\text{rep}}). \quad (36)$$

First, we investigate the sensitivity of the methods to the data-generating parameter values  $\theta$ . Then, we examine their sensitivity to error in the user’s specification  $\theta'$  of the input parameters. Here, we describe our analysis using the notation  $\tilde{\beta}(t; \theta, \theta')$  to refer to transmission rate estimates constructed with user input  $\theta'$ , from data generated by parameter values  $\theta$ .

**2.6.1 Sensitivity to data-generating parameters.** In §3.5, we consider the ideal situation in which the input  $\theta'$  corresponds exactly to the data-generating  $\theta$ . In this case, how sensitive is error in  $\tilde{\beta}(t; \theta, \theta')$  to  $\theta$ ? For example, is  $\beta(t)$  estimated more accurately for diseases with longer mean generation interval  $t_{\text{gen}}$ , etc.? To answer these questions, we perform the following steps on a grid of data-generating parameter values  $\theta$ :

1. Simulate 1000 reported incidence time series using  $\theta$ .
2. Create corresponding mock (constant) birth and natural mortality time series (cf. §2.4), specifying  $v'_c = v_c$  and  $\mu'_c = \mu_c$  in the input  $\theta'$ .
3. Estimate  $\beta(t)$  from the simulated data, specifying  $S'_0 = S_0, I'_0 = I_0, t'_{\text{gen}} = t_{\text{gen}}, p'_{\text{rep}} = p_{\text{rep}}$ , and  $t'_{\text{rep}} = t_{\text{rep}}$  in the input  $\theta'$ .
4. Compute the median RRMSE in the estimates  $\tilde{\beta}(t_k; \theta, \theta')$  (1000 estimates corresponding to 1000 simulations).

We repeat this analysis 6 times, corresponding to 2 methods of  $\beta(t)$  estimation (S or SI) and 3 methods of data simulation:



- *without* demographic stochasticity and *without* observation error (fixing  $p_{\text{rep}} = 1$ ,  $t_{\text{rep}} = 0$  weeks),
- *with* demographic stochasticity but *without* observation error (fixing  $p_{\text{rep}} = 1$ ,  $t_{\text{rep}} = 0$  weeks), or
- *with* demographic stochasticity and *with* observation error (fixing  $p_{\text{rep}} = 0.25$  unless sensitivity to  $p_{\text{rep}}$  is considered,  $t_{\text{rep}} = 2$  weeks).

Environmental stochasticity ( $\epsilon = 0.5$ ) is included in all simulations.

**2.6.2 Sensitivity to mis-specification of input parameters.** In §3.6, we fix the data-generating  $\theta$  and consider the more realistic situation in which components of the input  $\theta'$  differ from the corresponding components of  $\theta$  by a potentially large factor. In this case, how sensitive is error in  $\tilde{\beta}(t; \theta, \theta')$  to error in  $\theta'$ ? For example, how important is having an accurate estimate of  $t_{\text{gen}}$ , *etc.*? To answer these questions, we perform the following steps:

1. Simulate 1000 reported incidence time series using fixed data-generating parameter values  $\theta$ . (We assign the reference values listed in Table 1.)
2. For each point on a grid of input parameter values  $\theta'$ :
  - a. Create mock (constant) birth and natural mortality time series, taking  $v'_c$  and  $\mu'_c$  from the input  $\theta'$ .
  - b. Estimate  $\beta(t)$  from the simulated data, taking  $S'_0, I'_0, t'_{\text{gen}}, p'_{\text{rep}}$ , and  $t'_{\text{rep}}$  from the input  $\theta'$ .
  - c. Compute the median RRMSE in the estimates  $\tilde{\beta}(t_k; \theta, \theta')$  (1000 estimates corresponding to 1000 simulations).

We repeat this analysis 6 times, as outlined at the end of §2.6.1.

## 2.7 Asymptotic analysis

Here, we examine analytically the propagation of input error to the output of the SI method. (Similar expressions for propagated errors are obtained by analyzing the S method.) Our analysis here supports numerical results presented in §3.6 concerning the sensitivity of  $\beta(t)$  estimation error to mis-specification of input parameters.

**2.7.1 Explicit solutions of the  $(S_k, I_k)$  difference equations.** The SI method uses Eq (26a) to Eq (26c) to recursively reconstruct  $S(t)$  and  $I(t)$  from time series of reported incidence, births, and natural mortality. After substitution of Eq (26a), Eq (26b) and Eq (26c) can be written as

$$S_{k+1} = \frac{1 - \frac{1}{2}\mu_k\Delta t}{1 + \frac{1}{2}\mu_{k+1}\Delta t} S_k + \frac{B_{k+1} - \frac{1}{p_{\text{rep}}}C_{k+1+r}}{1 + \frac{1}{2}\mu_{k+1}\Delta t}, \quad k = 0, 1, \dots, \quad (37a)$$

$$I_{k+1} = \frac{1 - \frac{1}{2}(\gamma + \mu_k)\Delta t}{1 + \frac{1}{2}(\gamma + \mu_{k+1})\Delta t} I_k + \frac{\frac{1}{p_{\text{rep}}}C_{k+1+r}}{1 + \frac{1}{2}(\gamma + \mu_{k+1})\Delta t}, \quad k = 0, 1, \dots, \quad (37b)$$

where  $r = \lceil t_{\text{rep}} \rceil_{\Delta t} / \Delta t$  is the mean case reporting delay in units of the observation interval, rounded to the nearest integer. Eq (37) are linear, first order difference equations, whose explicit solutions are obtained using standard algebraic techniques (see Eq 1.2.4 in [41]) and

given by

$$S_k = S_0 \prod_{j=0}^{k-1} \frac{1 - \frac{1}{2} \mu_j \Delta t}{1 + \frac{1}{2} \mu_{j+1} \Delta t} + \sum_{i=0}^{k-1} (B_{i+1} - \frac{1}{p_{\text{rep}}} C_{i+1+r}) \prod_{j=i+1}^{k-1} \frac{1 - \frac{1}{2} \mu_j \Delta t}{1 + \frac{1}{2} \mu_{j+1} \Delta t}, \quad k = 0, 1, \dots, \quad (38a)$$

$$I_k = I_0 \prod_{j=0}^{k-1} \frac{1 - \frac{1}{2} (\gamma + \mu_j) \Delta t}{1 + \frac{1}{2} (\gamma + \mu_{j+1}) \Delta t} + \sum_{i=0}^{k-1} \frac{1}{p_{\text{rep}}} C_{i+1+r} \prod_{j=i+1}^{k-1} \frac{1 - \frac{1}{2} (\gamma + \mu_j) \Delta t}{1 + \frac{1}{2} (\gamma + \mu_{j+1}) \Delta t}, \quad k = 0, 1, \dots, \quad (38b)$$

with the conventions  $\prod_{i=b}^a x_i = 0$  and  $\prod_{i=b}^a x_i = 1$  if  $a < b$ . As we show in §2.7.2, explicit solutions of Eq (37) facilitate asymptotic analysis.

**2.7.2 Propagation of input error to  $(S_k, I_k)$ .** We consider the special case in which the vital rates are constant and set  $B_k = v_c \hat{N}_0 \Delta t$  and  $\mu_k = \mu_c$  for all  $k$  (cf. §2.4). Then Eq (38) simplify to

$$\begin{aligned} & S_k(S_0, v_c, \mu_c, p_{\text{rep}}) \\ &= S_0 \left( \frac{1 - \frac{1}{2} \mu_c \Delta t}{1 + \frac{1}{2} \mu_c \Delta t} \right)^k + \sum_{i=0}^{k-1} \frac{v_c \hat{N}_0 \Delta t - \frac{1}{p_{\text{rep}}} C_{i+1+r}}{1 + \frac{1}{2} \mu_c \Delta t} \left( \frac{1 - \frac{1}{2} \mu_c \Delta t}{1 + \frac{1}{2} \mu_c \Delta t} \right)^{k-1-i} \end{aligned} \quad (39a)$$

$$\begin{aligned} & I_k(I_0, \mu_c, t_{\text{gen}}, p_{\text{rep}}) \\ &= I_0 \left( \frac{1 - \frac{1}{2} (\gamma + \mu_c) \Delta t}{1 + \frac{1}{2} (\gamma + \mu_c) \Delta t} \right)^k + \sum_{i=0}^{k-1} \frac{\frac{1}{p_{\text{rep}}} C_{i+1+r}}{1 + \frac{1}{2} (\gamma + \mu_c) \Delta t} \left( \frac{1 - \frac{1}{2} (\gamma + \mu_c) \Delta t}{1 + \frac{1}{2} (\gamma + \mu_c) \Delta t} \right)^{k-1-i} \end{aligned} \quad (39b)$$

where we have made explicit the dependence of  $S_k$  and  $I_k$  on input parameters  $S_0, I_0, v_c, \mu_c, t_{\text{gen}} = \gamma^{-1}$ , and  $p_{\text{rep}}$ . Using Eq (39), we can derive exact expressions for the error propagated to  $S_k$  and  $I_k$  in the SI method as a result of assigning an incorrect value to an input parameter.

If the initial number of susceptibles is truly  $S_0$ , but we specify  $S'_0 = \omega S_0$ , where  $\omega > 0$ , then the error propagated to  $S_k$  is

$$\begin{aligned} \text{Err}(S_k, S_0 \leftarrow \omega S_0) &= S_k(\omega S_0, v_c, \mu_c, p_{\text{rep}}) - S_k(S_0, v_c, \mu_c, p_{\text{rep}}) \\ &= (\omega - 1) S_0 \left( \frac{1 - \frac{1}{2} \mu_c \Delta t}{1 + \frac{1}{2} \mu_c \Delta t} \right)^k \\ &= (\omega - 1) S_0 \left( \frac{1 - \frac{\Delta t}{2 t_{\text{life}}}}{1 + \frac{\Delta t}{2 t_{\text{life}}}} \right)^k \xrightarrow{k \rightarrow \infty} 0, \end{aligned} \quad (40)$$

where  $t_{\text{life}} = \mu_c^{-1}$  is the life expectancy in the population. Similarly, specifying  $I'_0 = \omega I_0$  for  $I_0$  yields an error

$$\begin{aligned}
 \text{Err}(I_k, I_0 \leftarrow \omega I_0) &= I_k(\omega I_0, \mu_c, t_{\text{gen}}, p_{\text{rep}}) - I_k(I_0, \mu_c, t_{\text{gen}}, p_{\text{rep}}) \\
 &= (\omega - 1)I_0 \left( \frac{1 - \frac{1}{2}(\gamma + \mu_c)\Delta t}{1 + \frac{1}{2}(\gamma + \mu_c)\Delta t} \right)^k \\
 &= (\omega - 1)I_0 \left( \frac{1 - \frac{\Delta t}{2t_{\text{inf}}}}{1 + \frac{\Delta t}{2t_{\text{inf}}}} \right)^k \xrightarrow{k \rightarrow \infty} 0
 \end{aligned}
 \tag{41}$$

in  $I_k$ , where  $t_{\text{inf}} = (\gamma + \mu_c)^{-1}$  is the mean time between infection and removal from the infected compartment, accounting for the possibility of natural death during infection. Eqs (40) and (41) show that the errors propagated to  $S_k$  and  $I_k$  vanish as  $k \rightarrow \infty$ ; we exploit this fact to improve susceptible reconstruction (cf. §2.8).

Mis-specifying  $v_c$  by assigning a value  $v'_c = \omega v_c$  creates an error in  $S_k$  that increases in magnitude over time and converges to a limit:

$$\begin{aligned}
 \text{Err}(S_k, v_c \leftarrow \omega v_c) &= S_k(S_0, \omega v_c, \mu_c, p_{\text{rep}}) - S_k(S_0, v_c, \mu_c, p_{\text{rep}}) \\
 &= \sum_{i=0}^{k-1} \frac{(\omega - 1)v_c \hat{N}_0 \Delta t}{1 + \frac{1}{2}\mu_c \Delta t} \left( \frac{1 - \frac{1}{2}\mu_c \Delta t}{1 + \frac{1}{2}\mu_c \Delta t} \right)^{k-1-i} \\
 &= \frac{(\omega - 1)v_c \hat{N}_0 \Delta t}{1 + \frac{1}{2}\mu_c \Delta t} \sum_{i=0}^{k-1} \left( \frac{1 - \frac{1}{2}\mu_c \Delta t}{1 + \frac{1}{2}\mu_c \Delta t} \right)^i \\
 &= \frac{(\omega - 1)v_c \hat{N}_0}{\mu_c} \left[ 1 - \left( \frac{1 - \frac{1}{2}\mu_c \Delta t}{1 + \frac{1}{2}\mu_c \Delta t} \right)^k \right] \xrightarrow{k \rightarrow \infty} (\omega - 1)v_c \hat{N}_0 t_{\text{life}}.
 \end{aligned}
 \tag{42}$$

Unlike Eq (42), the exact expression for  $\text{Err}(S_k, \mu_c \leftarrow \omega \mu_c)$  is not readily simplified and is difficult to interpret:

$$\begin{aligned}
 \text{Err}(S_k, \mu_c \leftarrow \omega \mu_c) &= S_k(S_0, v_c, \omega \mu_c, p_{\text{rep}}) - S_k(S_0, v_c, \mu_c, p_{\text{rep}}) \\
 &= S_0 \left( \frac{1 - \frac{1}{2}\omega \mu_c \Delta t}{1 + \frac{1}{2}\omega \mu_c \Delta t} \right)^k + \sum_{i=0}^{k-1} \frac{v_c \hat{N}_0 \Delta t - \frac{1}{p_{\text{rep}}} C_{i+1+r}}{1 + \frac{1}{2}\omega \mu_c \Delta t} \left( \frac{1 - \frac{1}{2}\omega \mu_c \Delta t}{1 + \frac{1}{2}\omega \mu_c \Delta t} \right)^{k-1-i} \\
 &\quad - S_0 \left( \frac{1 - \frac{1}{2}\mu_c \Delta t}{1 + \frac{1}{2}\mu_c \Delta t} \right)^k - \sum_{i=0}^{k-1} \frac{v_c \hat{N}_0 \Delta t - \frac{1}{p_{\text{rep}}} C_{i+1+r}}{1 + \frac{1}{2}\mu_c \Delta t} \left( \frac{1 - \frac{1}{2}\mu_c \Delta t}{1 + \frac{1}{2}\mu_c \Delta t} \right)^{k-1-i}.
 \end{aligned}
 \tag{43}$$

However, if  $C_k$  has a well-defined long-term average  $\langle C \rangle$  (this will be true if, for instance,  $C_k$  is periodic), then  $\text{Err}(S_k, \mu_c \leftarrow \omega \mu_c)$  has a well-defined long-term average  $\langle \text{Err}(S_k, \mu_c \leftarrow \omega \mu_c) \rangle$  with a simple form. Replacing  $C_{i+1+r}$  in Eq (43) with  $\langle C \rangle$ , simplifying the resulting expression,

then taking the limit as  $k \rightarrow \infty$ , we obtain

$$\begin{aligned} & \langle \text{Err}(S_k, \mu_c \leftarrow \omega \mu_c) \rangle \\ &= \lim_{k \rightarrow \infty} \left\{ S_0 \left( \frac{1 - \frac{1}{2} \omega \mu_c \Delta t}{1 + \frac{1}{2} \omega \mu_c \Delta t} \right)^k + \frac{v_c \hat{N}_0 \Delta t - \frac{1}{p_{\text{rep}}} \langle C \rangle}{\omega \mu_c \Delta t} \left[ 1 - \left( \frac{1 - \frac{1}{2} \omega \mu_c \Delta t}{1 + \frac{1}{2} \omega \mu_c \Delta t} \right)^k \right] \right. \\ & \quad \left. - S_0 \left( \frac{1 - \frac{1}{2} \mu_c \Delta t}{1 + \frac{1}{2} \mu_c \Delta t} \right)^k - \frac{v_c \hat{N}_0 \Delta t - \frac{1}{p_{\text{rep}}} \langle C \rangle}{\mu_c \Delta t} \left[ 1 - \left( \frac{1 - \frac{1}{2} \mu_c \Delta t}{1 + \frac{1}{2} \mu_c \Delta t} \right)^k \right] \right\} \\ &= \left( \frac{1}{\omega} - 1 \right) \left( v_c \hat{N}_0 \Delta t - \frac{1}{p_{\text{rep}}} \langle C \rangle \right) \frac{t_{\text{lifc}}}{\Delta t}, \end{aligned} \tag{44}$$

We can similarly show the following, still assuming that  $\langle C \rangle$  is well-defined:

$$\langle \text{Err}(I_k, \mu_c \leftarrow \omega \mu_c) \rangle = \frac{\langle C \rangle (t'_{\text{inf}} - t_{\text{inf}})}{p_{\text{rep}} \Delta t}, \quad t'_{\text{inf}} = (\gamma + \omega \mu_c)^{-1}, \tag{45}$$

$$\langle \text{Err}(I_k, t_{\text{gen}} \leftarrow \omega t_{\text{gen}}) \rangle = \frac{\langle C \rangle (t'_{\text{inf}} - t_{\text{inf}})}{p_{\text{rep}} \Delta t}, \quad t'_{\text{inf}} = \left( \frac{1}{\omega} \gamma + \mu_c \right)^{-1}, \tag{46}$$

$$\langle \text{Err}(S_k, p_{\text{rep}} \leftarrow \omega p_{\text{rep}}) \rangle = \left( 1 - \frac{1}{\omega} \right) \frac{\langle C \rangle t_{\text{lifc}}}{p_{\text{rep}} \Delta t}, \tag{47}$$

$$\langle \text{Err}(I_k, p_{\text{rep}} \leftarrow \omega p_{\text{rep}}) \rangle = \left( \frac{1}{\omega} - 1 \right) \frac{\langle C \rangle t_{\text{inf}}}{p_{\text{rep}} \Delta t}. \tag{48}$$

Here,  $t'_{\text{inf}}$  is the (incorrect) mean time spent infected that results when  $\omega \mu_c$  is incorrectly specified for  $\mu_c$  (Eq 45) or  $\omega t_{\text{gen}}$  is incorrectly specified for  $t_{\text{gen}}$  (Eq 46).

**2.7.3 Propagation of error in  $(S_k, I_k)$  to  $\beta_k$ .** Let  $\beta_k(Z_k, Z_{k+1}, S_k, I_k)$  be the raw SI method estimate of  $\beta(t_k)$ , given by the right hand side of Eq (26d). Suppose that, due to propagated error (cf. §2.7.2), the arguments are incorrect by a factor, so that

$$Z_k = \omega_Z Z(t_k), \quad Z_{k+1} = \omega_Z Z(t_{k+1}), \quad S_k = \omega_S S(t_k), \quad I_k = \omega_I I(t_k), \tag{49}$$

where  $\omega_Z, \omega_S, \omega_I > 0$ . Then the computed  $\beta_k$  will have relative error

$$\frac{\beta_k(Z_k, Z_{k+1}, S_k, I_k) - \beta_k(Z(t_k), Z(t_{k+1}), S(t_k), I(t_k))}{\beta_k(Z(t_k), Z(t_{k+1}), S(t_k), I(t_k))} = \frac{\omega_Z}{\omega_S \omega_I} - 1. \tag{50}$$

Hence severe underestimation of  $S_k$  or  $I_k$  ( $\omega_S \ll 1$  or  $\omega_I \ll 1$ ) causes the relative error in  $\beta_k$  to blow up.

### 2.8 Estimating $S_0$ via peak-to-peak iteration

Reconstruction of susceptibles  $S(t)$  is a necessary step in the reconstruction of  $\beta(t)$  using the FC, S, and SI methods. In §3.6, we show that susceptible reconstruction requires accurate specification of the initial number of susceptibles  $S_0 = S(t_0)$ . However, reliable estimates of  $S_0$  have, to this point, been difficult to obtain in practice.

We propose a technique for iteratively improving estimates of  $S_0$ , requiring only incidence, birth, and natural mortality data at times  $t_k$  (Eq (3)). Crucially, our technique, which we call

“peak-to-peak iteration” (PTPI), enables accurate susceptible reconstruction without direct observation of the susceptible population size at the initial time.

Our approach is motivated by application of the SI method to simulated data. When we incorrectly guessed the value of  $S_0$  and attempted to reconstruct  $S(t)$  via Eq (26b), the absolute error in the reconstruction  $\{(t_k, S_k)\}_{k=0}^n$  decreased monotonically over time ( $k$ ). (Eq (40) shows that the error propagated from  $S_0$  to  $S_k$  vanishes as  $k \rightarrow \infty$ .) Consequently, if the underlying dynamics are at least approximately periodic, and if  $t_0$  and  $t_n$  occur at the same phase of the cycle, then  $S_n$  is actually a better estimate of  $S_0$  than our initial guess. In this situation, instead of reconstructing  $\beta(t)$  directly, we can use  $S_n$  as an updated estimate of  $S_0$ , and reconstruct  $S(t)$  more accurately. This procedure can be repeated any number of times, and, with simulated data, we observe rapid convergence to an accurate estimate of  $S_0$  (cf. §3.7).

The key point is that the reconstructed final state can be used as an improved estimate of the initial state only if the initial and final states occur at the same phase of the cycle. This will not be true unless the observation period (the time between the first and last observations in time series data) is an integer multiple of the period of the underlying dynamics. We can ensure this by choosing appropriate times at which to start and stop  $S(t)$  reconstruction. In noisy periodic data, the points in a cycle that are easiest to identify robustly are the peaks. Consequently, we ignore observations (i) prior to the time  $t_a$  of the first peak in the incidence time series and (ii) after the time  $t_b$  of the last peak that occurs near an integer multiple of the apparent period after the first peak. For the truncated time series, the iterations converge to an accurate estimate of  $S(t_a)$  starting from an initial guess, and we recover the corresponding accurate estimate of  $S_0$  by solving Eq (26b) backwards in time, from  $t_a$  to  $t_0$ :

$$S_{k-1} \approx \frac{[1 + \frac{1}{2}\mu_k\Delta t]S_k - B_k + Z_k}{1 - \frac{1}{2}\mu_{k-1}\Delta t}. \tag{51}$$

The complete PTPI algorithm, which consists of finding  $t_a$  and  $t_b$  (truncation step) and estimating  $S_0$  (iteration step), is outlined in Boxes 5 and 6 below. In §3.7, we assess the performance of PTPI by applying the technique to simulated data with known underlying  $S_0$ , starting from an incorrect initial estimate of  $S_0$ .

### Box 5. Peak-to-peak iteration: Truncation step

**Goal:** Given a roughly periodic time series  $\{(t_k, Z_k)\}_{k=0}^n$  of incidence, we want to find the time  $t_a$  of the first peak and the time  $t_b$  of the last peak occurring at the same phase of the cycle. These times are necessary for the iteration step (Box 6).

**Algorithm:**

- i. Smooth the raw incidence time series  $Z_k$  by applying a  $(2\ell_1 + 1)$ -point central moving average, computed via

$$\bar{Z}_k = \frac{1}{2\ell_1 + 1} \sum_{i=-\ell_1}^{\ell_1} Z_{k+i}, \quad k = \ell_1, \dots, n - \ell_1. \tag{52}$$

Choose minimal  $\ell_1$  large enough to remove spurious peaks in  $Z_k$  caused by noise, while retaining true peaks.

- ii. Identify the period  $T$  of the smoothed incidence time series  $\{(t_k, \bar{Z}_k)\}_{k=\ell_1}^{n-\ell_1}$  from its power spectrum, and calculate the number of embedded cycles, given by  $m = \lfloor \frac{t_{n-\ell_1} - t_{\ell_1}}{T} \rfloor$ .

iii. Construct the set  $\mathcal{I}$  indexing peaks in  $\{(t_k, \bar{Z}_k)\}_{k=\ell_1}^{n-\ell_2}$ :

$$\mathcal{I} = \{k \in \{\ell_1 + \ell_2, \dots, n - \ell_1 - \ell_2\} : \bar{Z}_k > \bar{Z}_{k \pm i} \text{ for all } i = 1, \dots, \ell_2\}. \quad (53)$$

Choose minimal  $\ell_2$  large enough to ensure that  $\mathcal{I}$  indexes true peaks in  $\bar{Z}_k$ , but not spurious peaks caused by noise (any that remain after smoothing).

iv. Define  $\mathcal{T} = \{t_k : k \in \mathcal{I}\}$ , the set of times of peaks in  $\bar{Z}_k$ , and record the time of the first peak, given by  $t_a = \min(\mathcal{T})$ .

v. For  $i = 0, \dots, m$ , define  $\tau_i = t_a + iT$  and find the element of  $\mathcal{T}$  nearest  $\tau_i$ , namely  $\arg \min_{\tau \in \mathcal{T}} |\tau_i - \tau|$ . The resulting subset  $\mathcal{T}_{\text{phase}}$  should contain successive time points that are roughly one period apart, *i.e.*, the corresponding peaks in  $\bar{Z}_k$  should occur at the same phase of the cycle.

vi. Record the time of the last such peak, given by  $t_b = \max(\mathcal{T}_{\text{phase}})$ .

### Box 6. Peak-to-peak iteration: Iteration step

**Goal:** We want to produce an accurate estimate of the initial number of susceptibles  $S_0 = S(t_0)$ , given

- a roughly periodic time series  $\{(t_k, Z_k)\}_{k=0}^n$  of incidence,
- a time series  $\{(t_k, B_k)\}_{k=0}^n$  of births,
- a time series  $\{(t_k, \mu_k)\}_{k=0}^n$  of the *per capita* natural mortality rate,
- times  $t_a$  and  $t_b$  as defined in the truncation step (Box 5), and
- an initial estimate of  $S_0$ .

#### Algorithm:

- Define an initial estimate of  $S(t_a)$ . (We use the initial estimate of  $S_0$ .)
- Reconstruct  $S(t)$  between times  $t_a$  and  $t_b$  using Eq (26b), starting with the current estimate of  $S(t_a)$ .
- Update the estimate of  $S(t_a)$  with the estimate of  $S(t_b)$  obtained in (ii).
- Repeat (ii) and (iii) until the sequence of estimates of  $S(t_a)$  converges (to within a desirable tolerance).
- Reconstruct  $S(t)$  between times  $t_0$  and  $t_a$  using Eq (51), starting with the final estimate of  $S(t_a)$  obtained in (iv). The reconstruction is performed backwards in time, from  $t_a$  to  $t_0$ .
- Record the estimate of  $S_0 = S(t_0)$  computed in (v). This value can be passed back to Eq (26b), allowing for reconstruction of  $S(t)$  between times  $t_0$  and  $t_m$  as usual.

### 3 Results

In §3.1, we compare the performance of the FC, S, and SI methods in estimating  $\beta(t)$  from an idealized reported incidence time series. In §3.2, we show how process and observation error create spurious noise in estimates of  $\beta(t)$ . In §§3.3 and 3.4, we examine averaging and smoothing as ways to distill temporal patterns of interest from noisy estimates of  $\beta(t)$ . In §§3.5 and 3.6, we summarize our systematic analysis of the sensitivity of  $\beta(t)$  estimation error to data-generating parameters and to mis-specification of input parameters by the user. In §3.7, addressing apparent sensitivity to mis-specification of the initial number of susceptibles  $S_0$ , we assess the performance of PTPI as a method of estimating  $S_0$ . Finally, in §3.8, we report the run times of the S and SI methods and PTPI.

The results reported here are entirely reproducible using the annotated R code available in [S1 File](#).

#### 3.1 Example of $\beta(t)$ estimation using the FC, S, and SI methods

We applied the FC, S, and SI methods without input error to estimate  $S(t)$  and  $\beta(t)$  from an idealized reported incidence time series, simulated without process or observation error. The time series estimates  $S_k$  and  $\beta_k$  are shown in [Fig 1](#). The S and SI methods estimated  $S(t)$  and  $\beta(t)$  accurately at every time point, whereas the FC method captured seasonality but failed otherwise. In the FC method,  $S_k$  neglects natural mortality ([Eq \(24b\)](#)), so it increases without bound while  $\beta_k$  decays to zero due to division by  $S_k$  ([Eq \(24d\)](#)).

[Fig 1A](#) confirms that the absolute error in the FC method estimate of  $S(t)$  increases linearly as  $\mu_c \langle S \rangle t$ , where  $\mu_c$  is the constant *per capita* natural mortality rate and  $\langle S \rangle$  is the continuous-time average of  $S(t)$ . In practice, the FC method fails whenever natural mortality in the underlying population is non-negligible. Since the S and SI methods address this limitation at effectively no computational cost, we do not present further analysis of the FC method.

In [Fig 1B](#), the SI method estimate of  $\beta(t)$  was very accurate (RRMSE  $\approx 0.2\%$ ), whereas the S method estimate peaked too early and too high (RRMSE  $\approx 2.4\%$ ).

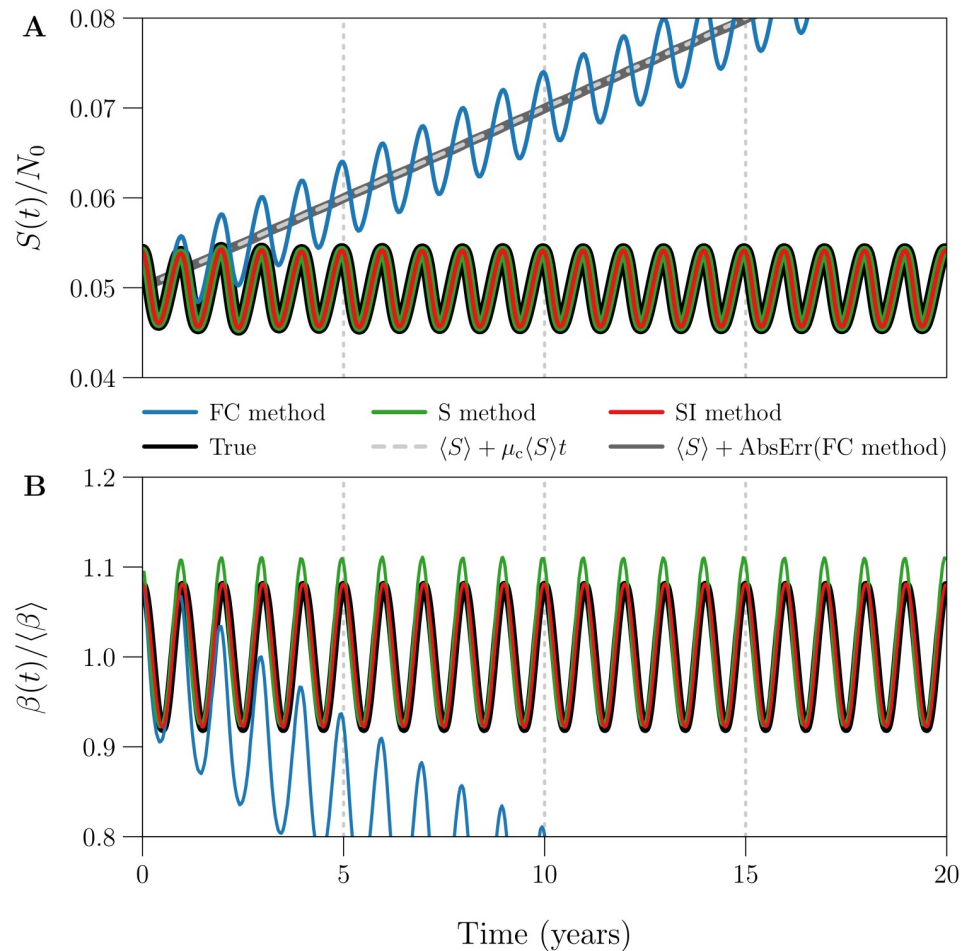
#### 3.2 Effects of process and observation error

We applied the S and SI methods without input error to four reported incidence time series  $C_k$ , simulated using the same parameter values but with different levels of process and observation noise. The first simulation was purely deterministic, while the remaining three included (i) environmental stochasticity [ES], (ii) ES and demographic stochasticity [ES+DS], or (iii) ES, DS, and observation error [ES+DS+OE]. [Fig 2](#) shows the resulting estimates  $Z_k$ ,  $I_k$ , and  $\beta_k$  of true incidence  $Z(t)$ , prevalence  $I(t)$ , and the seasonally forced transmission rate  $\beta(t)$ .

Noise of any type introduces random fluctuations in  $C_k$  on top of longer-term (e.g., seasonal) variation. Noise in  $C_k$  is propagated to  $Z_k$  ([Fig 2A](#)) and  $I_k$  ([Fig 2B](#)), because (i) in both the S and SI methods, we scale  $C_{k+r}$  by a constant factor of  $p_{\text{rep}}^{-1} \geq 1$  to compute  $Z_k$  (Eqs (25a) and (26a)); (ii) in the S method, we scale  $C_{k+1-g+r}$  by a constant factor of  $[p_{\text{rep}}(\gamma + \mu_k)\Delta t]^{-1}$  to compute  $I_k$  ([Eq \(25c\)](#) after substitution of [Eq \(25a\)](#)); and (iii) in the SI method,  $I_k$  contains a weighted sum of  $C_i$  terms ([Eq \(38b\)](#)).

Noise in  $Z_k$  and  $I_k$  is amplified in  $\beta_k$  ([Fig 2C](#)), distorting the correct temporal pattern, for the following reason. When  $Z$  and  $I$  are close to zero, small absolute changes in either yield large relative changes in the ratio  $Z/I$  and in turn  $\beta_k$ , which contains a factor of  $Z_{k+1}/I_k$  in the S method ([Eq \(25d\)](#)) and  $(Z_k + Z_{k+1})/(2I_k)$  in the SI method ([Eq \(26d\)](#)). Hence low amplitude



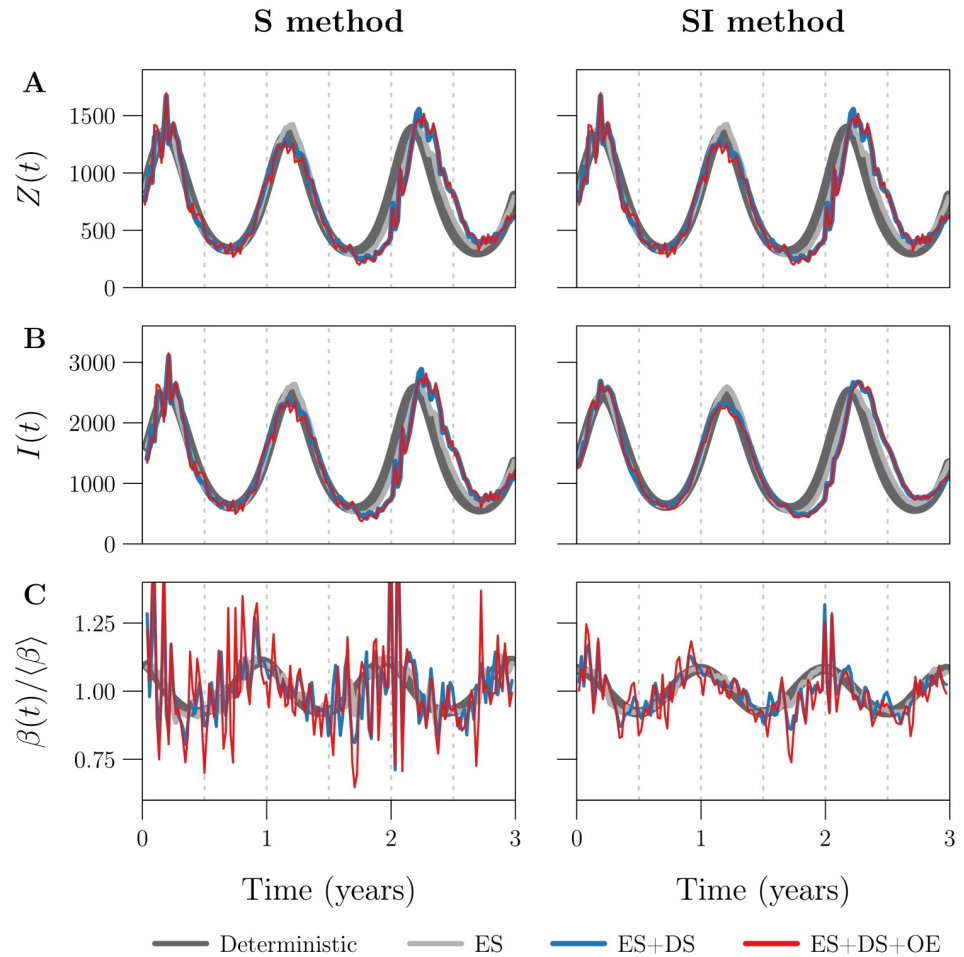


**Fig 1. Example of  $S(t)$  and  $\beta(t)$  estimation using the FC, S, and SI methods.** Plotted are the susceptible population size  $S(t)$  and seasonally forced transmission rate  $\beta(t)$  (Eq (27)) underlying 20 years of weekly reported incidence, together with time series estimates  $S_k$  and  $\beta_k$  obtained from the data by the FC [blue], S [green], and SI [red] methods. The reported incidence time series ( $\Delta t = 1$  week,  $n = \lfloor 20 \times 365/7 \rfloor = 1042$ ) was simulated without process or observation error ( $\epsilon = 0$ ,  $p_{\text{rep}} = 1$ ), using reference values (Table 1) for all other data-generating parameters. The three estimation methods were applied without input error, *i.e.*, all input parameters were assigned their true (data-generating) values. **[Panel A]**  $S(t)$  scaled by  $1/N_0$ , describing the number of susceptibles as a proportion of the initial population size. Grey lines show that the absolute error in the FC method estimate of  $S(t)$  increases linearly as  $\mu_c \langle S \rangle t$ , where  $\mu_c$  is the constant *per capita* natural mortality rate and  $\langle S \rangle$  is the continuous-time average of  $S(t)$ . **[Panel B]**  $\beta(t)$  scaled by  $1/\langle \beta \rangle$ , describing the transmission rate relative to its mean. RRMSE (Eq (33)) in the  $\beta_k$  time series generated by the (FC, S, SI) method is roughly (0.3355, 0.0240, 0.0021).

<https://doi.org/10.1371/journal.pcbi.1008124.g001>

noise in  $Z_k$  and  $I_k$  appears as spurious, higher amplitude noise in  $\beta_k$ . This is an important issue in practice, because the incidence of endemic diseases is typically very small relative to the population size, and periodic fluctuations bringing incidence even closer to zero are common for many diseases [4, 14, 42].

Fig 2 shows that the SI method is much better than the S method at resisting noise propagation. One reason is the effective smoothing of incidence in the SI method, which replaces  $Z_{k+1}$  with  $(Z_k + Z_{k+1})/2$  in the computation of  $\beta_k$  (compare Eqs (25d) and (26d)). We expose a second reason in §3.2.1 below by comparing the variance in  $I_k$  induced by observation error, between the two methods. (We expect similar results for process error.)



**Fig 2. Effects of process and observation error on the S and SI methods.** Plotted are the estimates [Row A]  $Z_k$ , [Row B]  $I_k$ , and [Row C]  $\beta_k$  of true incidence  $Z(t)$ , prevalence  $I(t)$ , and the seasonally forced transmission rate  $\beta(t)$  (Eq (27)) obtained by applying the [Left] S and [Right] SI methods without input error to each of four simulated reported incidence time series (indicated by the legend;  $\Delta t = 1$  week,  $n = \lfloor 3 \times 365/7 \rfloor = 156$ ). The first simulation was purely deterministic [dark grey] ( $\epsilon = 0, p_{\text{rep}} = 1$ ), while the remaining three accounted for (i) environmental stochasticity [ES, light grey] ( $\epsilon = 0.5, p_{\text{rep}} = 1$ ), (ii) ES and demographic stochasticity [ES+DS, blue] ( $\epsilon = 0.5, p_{\text{rep}} = 1$ ), or (iii) ES, DS, and observation error [ES+DS+OE, red] ( $\epsilon = 0.5, p_{\text{rep}} = 0.25$ ). Reference values (Table 1) were assigned to all other data-generating parameters, in all four simulations. The left and right panels in Row A are identical, because the S and SI methods compute  $Z_k$  identically (compare Eqs (25a) and (26a)). RRMSE in the  $\beta_k$  time series is (0.0239, 0.0375, 0.1126, 0.1432) with the S method and (0.0021, 0.0153, 0.0494, 0.0591) with the SI method (order follows the legend). Note that the underlying  $\beta(t)$  was the same in all simulations; it is not plotted in Row C, but is close to perfectly represented by the dark grey curve in the right panel (RRMSE  $\approx 0.2\%$ ). Due to process error, the underlying  $Z(t)$  and  $I(t)$  (also not shown) varied between the deterministic, ES, and ES+DS simulations.

<https://doi.org/10.1371/journal.pcbi.1008124.g002>

**3.2.1 Propagation of noise from  $C_k$  to  $I_k$ .** Consider the S and SI method estimates of prevalence  $I(t_k)$ ,

$$I_k^{[S]} = \frac{C_{k+1-g+r}}{p_{\text{rep}}(\gamma + \mu_c)\Delta t}, \tag{54a}$$

$$I_k^{[SI]} = I_0 \left( \frac{1 - \frac{1}{2}(\gamma + \mu_c)\Delta t}{1 + \frac{1}{2}(\gamma + \mu_c)\Delta t} \right)^k + \sum_{i=0}^{k-1} \frac{C_{i+1+r}}{p_{\text{rep}}[1 + \frac{1}{2}(\gamma + \mu_c)\Delta t]} \left( \frac{1 - \frac{1}{2}(\gamma + \mu_c)\Delta t}{1 + \frac{1}{2}(\gamma + \mu_c)\Delta t} \right)^{k-1-i}. \tag{54b}$$

Here,  $g = \lceil t_{\text{gen}} \rceil \Delta t / \Delta t$  and  $r = \lceil t_{\text{rep}} \rceil \Delta t / \Delta t$  are the mean generation interval and case reporting delay in units of the observation interval, rounded to the nearest integer. These estimates are obtained from Eq (25c) (after substitution of Eqs (25a) and (38b) when we assume a constant natural mortality rate  $\mu_c$ ). Following §2.3.3, suppose reported incidence is generated from true incidence  $Z(t_k)$  via  $C_{k+r} \stackrel{\text{ind.}}{\sim} \text{Binomial}(Z(t_k), p_{\text{rep}})$ . Then the variance of  $C_{k+r}$  is

$$\text{Var}(C_{k+r}) = Z(t_k)p_{\text{rep}}(1 - p_{\text{rep}}). \tag{55}$$

It follows from Eqs (54) and (55) and the identity  $\text{Var}(aX) = a^2 \text{Var}(X)$  that

$$\text{Var}(I_k^{[S]}) = \frac{(1 - p_{\text{rep}})Z(t_{k+1-g})}{p_{\text{rep}}[(\gamma + \mu_c)\Delta t]^2}, \tag{56a}$$

$$\text{Var}(I_k^{[SI]}) = \sum_{i=0}^{k-1} \frac{(1 - p_{\text{rep}})Z(t_{i+1})}{p_{\text{rep}}[1 + \frac{1}{2}(\gamma + \mu_c)\Delta t]^2} \left( \frac{1 - \frac{1}{2}(\gamma + \mu_c)\Delta t}{1 + \frac{1}{2}(\gamma + \mu_c)\Delta t} \right)^{2(k-1-i)}. \tag{56b}$$

If  $Z(t)$  has a well-defined average  $\langle Z \rangle$ , then replacing instances of  $Z$  in Eq (56) with  $\langle Z \rangle$  and taking the limit as  $k \rightarrow \infty$ , we obtain the average variances

$$\langle \text{Var}(I_k^{[S]}) \rangle = \frac{(1 - p_{\text{rep}})\langle Z \rangle}{p_{\text{rep}}[(\gamma + \mu_c)\Delta t]^2}, \tag{57a}$$

$$\begin{aligned} \langle \text{Var}(I_k^{[SI]}) \rangle &= \lim_{k \rightarrow \infty} \left\{ \frac{(1 - p_{\text{rep}})\langle Z \rangle}{p_{\text{rep}}[1 + \frac{1}{2}(\gamma + \mu_c)\Delta t]^2} \sum_{i=0}^{k-1} \left( \frac{1 - \frac{1}{2}(\gamma + \mu_c)\Delta t}{1 + \frac{1}{2}(\gamma + \mu_c)\Delta t} \right)^{2i} \right\} \\ &= \lim_{k \rightarrow \infty} \left\{ \frac{(1 - p_{\text{rep}})\langle Z \rangle}{2p_{\text{rep}}(\gamma + \mu_c)\Delta t} \left[ 1 - \left( \frac{1 - \frac{1}{2}(\gamma + \mu_c)\Delta t}{1 + \frac{1}{2}(\gamma + \mu_c)\Delta t} \right)^{2k} \right] \right\} \\ &= \frac{(1 - p_{\text{rep}})\langle Z \rangle}{2p_{\text{rep}}(\gamma + \mu_c)\Delta t}. \end{aligned} \tag{57b}$$

Comparing these with  $\langle \text{Var}(C_k) \rangle = \langle Z \rangle p_{\text{rep}}(1 - p_{\text{rep}})$  using reference parameter values  $t_{\text{gen}} = \gamma^{-1} = 13$  days,  $\mu_c = 0.04 \text{ year}^{-1}$ , and  $\Delta t = 1$  week, we obtain

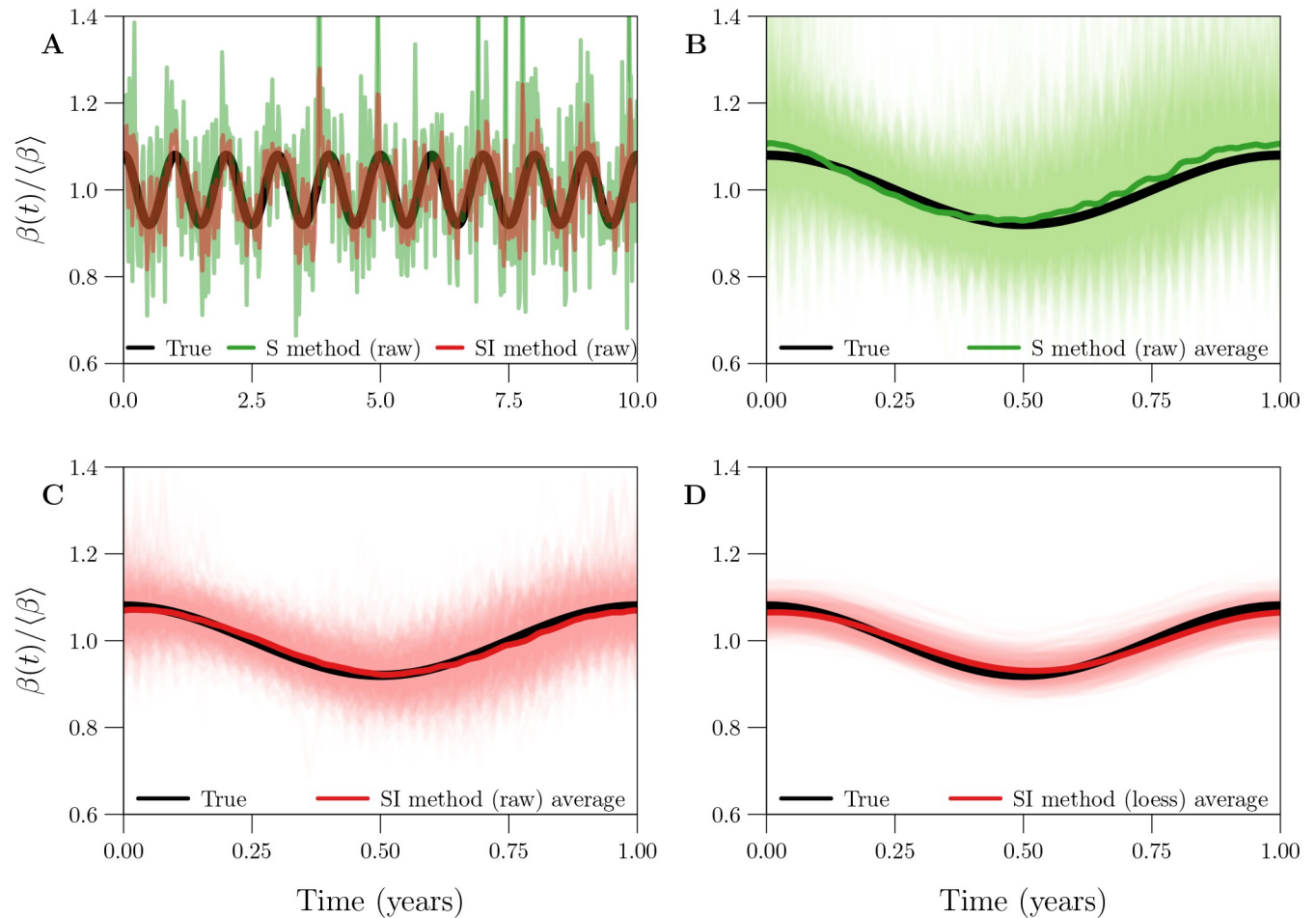
$$\frac{\langle \text{Var}(I_k^{[S]}) \rangle}{\langle \text{Var}(C_k) \rangle} = \frac{1}{p_{\text{rep}}^2[(\gamma + \mu_c)\Delta t]^2} = \frac{1}{p_{\text{rep}}^2} \left( \frac{t_{\text{inf}}}{\Delta t} \right)^2 \approx \frac{3.44}{p_{\text{rep}}^2}, \tag{58a}$$

$$\frac{\langle \text{Var}(I_k^{[SI]}) \rangle}{\langle \text{Var}(C_k) \rangle} = \frac{1}{2p_{\text{rep}}^2[(\gamma + \mu_c)\Delta t]^2} = \frac{1}{2p_{\text{rep}}^2} \left( \frac{t_{\text{inf}}}{\Delta t} \right)^2 \approx \frac{0.93}{p_{\text{rep}}^2}, \tag{58b}$$

where  $t_{\text{inf}} = (\gamma + \mu_c)^{-1}$  is the mean time spent infected. Hence, while both the S and SI methods suffer from propagation of noise from reported incidence  $C_k$  to estimated prevalence  $I_k$ , particularly for  $p_{\text{rep}} \ll 1$ , the S method tends to be much worse (by a factor of  $3.44/0.93 \approx 3.7$  in this example). Comparative resistance to noise propagation is a distinct advantage of the SI method over the S method.

### 3.3 Averaging the raw estimate of $\beta(t)$

Fig 3A displays two raw estimates  $\beta_k$  (S and SI methods, applied without input error) of a seasonally forced  $\beta(t)$ , each spanning 1000 years (only the first 10 years are shown). The estimates



**Fig 3. Bias and variance in 1-year cycles embedded in three estimates of a seasonally forced  $\beta(t)$ .** [Panel A] In black, the seasonally forced  $\beta(t)$  (Eq (27)) underlying 1000 years of simulated reported incidence data. In (transparent) colour, raw estimates  $\beta_k$  obtained from the data by the S [green] and SI [red] methods, both applied without input error. Only the first 10 of 1000 years are shown. [Panels B and C] In black, the true 1-year cycle in the seasonally forced  $\beta(t)$ . In light (transparent) colour, the 1000 1-year cycles embedded in the linear interpolant  $\beta_{\text{int}}(t)$  of  $\beta_k$ . In dark colour, the average 1-year cycle (Eq (22a)) in  $\beta_{\text{int}}(t)$ . Results are shown for both the S [Panel B, green] and SI [Panel C, red] methods. [Panel D] Like Panel C, except for a smooth loess curve  $\beta_{\text{loess}}(t; q)$  ( $q = 53$ ) fit to  $\beta_k$ , instead of the interpolant  $\beta_{\text{int}}(t)$ . [Details] A reported incidence time series with 1000 years of weekly observations ( $\Delta t = 1$  week,  $n = 52153$ ) was simulated with environmental noise in transmission ( $\epsilon = 0.5$ ), demographic stochasticity, and random under-reporting of cases ( $p_{\text{rep}} = 0.25$ ), using reference values (Table 1) for the remaining parameters.

<https://doi.org/10.1371/journal.pcbi.1008124.g003>

embed 1000 1-year cycles, which are displayed in Fig 3B and 3C together with their 1-year average (cf. §2.2.5).

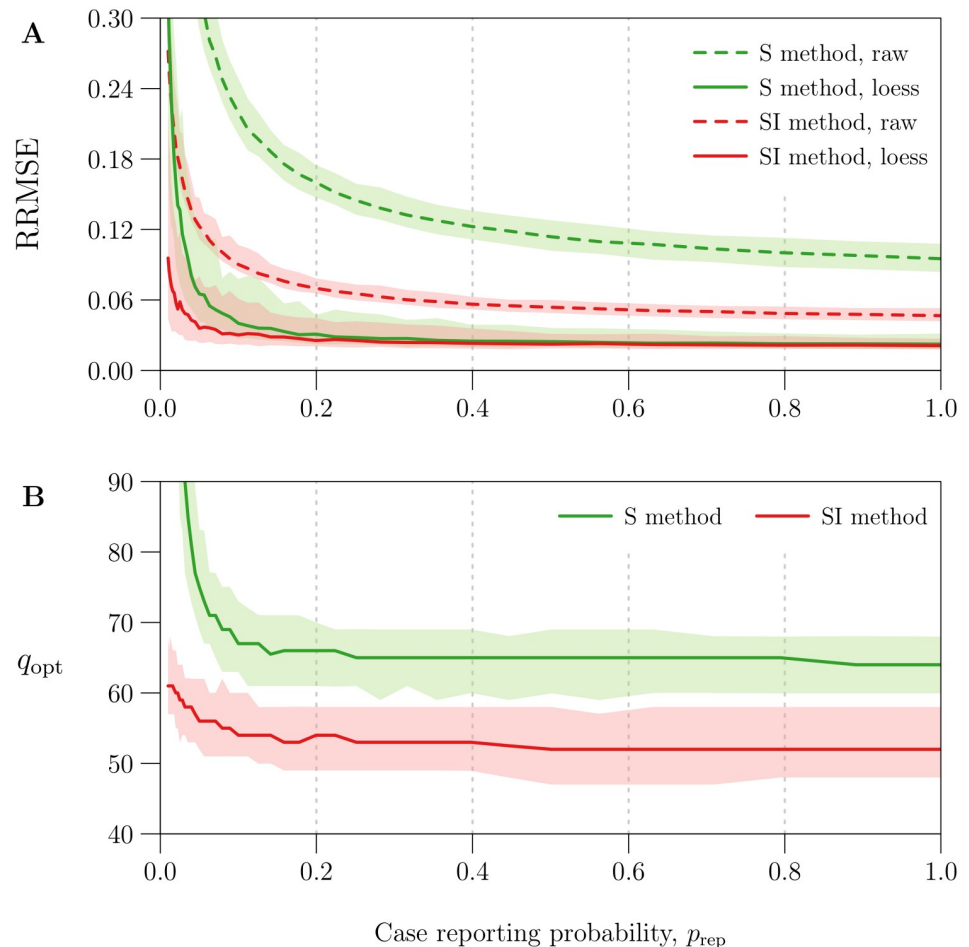
Both estimates suffered from spurious noise distorting the correct seasonal pattern, caused by process and observation error in the data-generating process (cf. §3.2). As in Fig 2C, the variance was markedly smaller with the SI method. Averaging the embedded 1-year cycles recovered the true 1-year cycle from the noise. In the absence of input error, the S method appears to carry a slight bias (peaking early and too high, as in Fig 1), whereas the SI method is nearly unbiased.

While some existing infectious disease time series span several centuries [15], in practice, averaging as in Fig 3B and 3C is sensible only over time intervals during which the underlying seasonal pattern in transmission is roughly stationary.

### 3.4 Smoothing the raw estimate of $\beta(t)$

Regardless of whether averaging is employed, comparison of Fig 3C and 3D shows that it is helpful to smooth the  $\beta_k$  time series by fitting a loess curve  $\beta_{\text{loess}}(t; q)$  (cf. §2.2.6). An appropriate degree of smoothing (*i.e.*, choice of loess smoothing parameter  $q$ ) eliminated spurious noise without significantly increasing bias.

Fig 4A quantifies the effect of smoothing  $\beta_k$  using the optimal value  $q_{\text{opt}}$  for parameter  $q$  (cf. §2.2.6). It plots RRMSE before and after smoothing as a function of the amount of noise in the simulated reported incidence data, which was modulated by varying the case reporting probability  $p_{\text{rep}}$  between 0.01 and 1 (more noise for smaller  $p_{\text{rep}}$ ; see Eq (31)).



**Fig 4. Reduction in  $\beta(t)$  estimation error with optimal loess smoothing.** The horizontal axis measures the case reporting probability  $p_{\text{rep}}$ , for which 41 values equally spaced on a logarithmic scale between 0.01 and 1 were considered. Using each value of  $p_{\text{rep}}$  and reference values (Table 1) for all other parameters, 100 reported incidence time series ( $\Delta t = 1$  week,  $n = 1042$ ) were simulated accounting for environmental noise in transmission ( $\epsilon = 0.5$ ), demographic stochasticity, and random under-reporting of cases (measured by  $p_{\text{rep}}$ ). The underlying seasonally forced  $\beta(t)$  (Eq (27)) was estimated from reported incidence using the S and SI methods, both applied without input error, yielding two raw estimates  $\beta_k$  per simulation. Smooth loess curves  $\beta_{\text{loess}}(t; q)$  ( $q = 10, \dots, 110$ ; cf. §2.2.6) were fit to each  $\beta_k$  time series. The optimal  $q$  for a given time series, denoted by  $q_{\text{opt}}$ , was defined as the value that minimized RRMSE (Eq (33)) in  $\beta_{\text{loess}}(t_k; q)$ . Overall, for each value of  $p_{\text{rep}}$  and each  $\beta(t)$  estimation method (S and SI), 100 values of  $q_{\text{opt}}$  were obtained corresponding to 100  $\beta_k$  time series. Plotted on the vertical axis as functions of  $p_{\text{rep}}$  are the median and 5th and 95th percentiles of [Panel A] RRMSE in the raw estimates  $\beta_k$  [dashed lines] and optimal loess estimates  $\beta_{\text{loess}}(t_k; q_{\text{opt}})$  [solid lines] and [Panel B]  $q_{\text{opt}}$ . Lines and bands indicate the median and 5th–95th percentile range, respectively. Results for the S and SI methods are shown in green and red, respectively.

<https://doi.org/10.1371/journal.pcbi.1008124.g004>



Using the optimal loess estimate  $\beta_{\text{loess}}(t_k; q_{\text{opt}})$  instead of the raw estimate  $\beta_k$  significantly reduced RRMSE—by at least 46% for the S method and 17% for the SI method across all simulations. Although raw estimates generated by the SI method were consistently more accurate (expected in light of Fig 3B and 3C), optimal loess estimates were comparable between the S and SI methods for  $p_{\text{rep}} > 0.2$  (RRMSE  $\approx$  3%). For  $p_{\text{rep}} < 0.2$  (severe under-reporting of cases), optimal smoothing failed to an increasing extent to recover the underlying  $\beta(t)$  from noise in  $\beta_k$ . In this setting, the S method was greatly outperformed by the SI method, which is more resilient to noise in reported incidence (cf. §3.2).

Fig 4B shows that median  $q_{\text{opt}}$  was roughly constant for  $p_{\text{rep}} > 0.1$ , with

$$\text{median } q_{\text{opt}} \approx \begin{cases} 65 & \text{for the S method,} \\ 53 & \text{for the SI method.} \end{cases} \quad (59)$$

More smoothing (greater  $q$ ) was required to minimize RRMSE for  $p_{\text{rep}} > 0.1$ . More generally, Fig 4 indicates that the S and SI methods should always include a smoothing step. Hence, in the remaining analysis, we always smooth  $\beta_k$ .

### 3.5 Sensitivity to data-generating parameters

Here, we characterize the sensitivity of  $\beta(t)$  estimation error to parameters of the data-generating process. As in §§3.1–3.4, we consider the ideal case in which the user-specified values of all input parameters are equal to the true (data-generating) values. The details of our analysis are outlined in §2.6.1.

Fig 5 plots the median RRMSE in estimates of a seasonally forced  $\beta(t)$  (Eq (27)) from 1000 realizations of a reported incidence time series, as a bivariate function of the mean  $\langle\beta\rangle$  and amplitude  $\alpha$  of seasonal forcing. To aid interpretation, the  $\langle\beta\rangle$  axis was scaled to measure the basic reproduction number  $\mathcal{R}_0$  (Eq (2)).

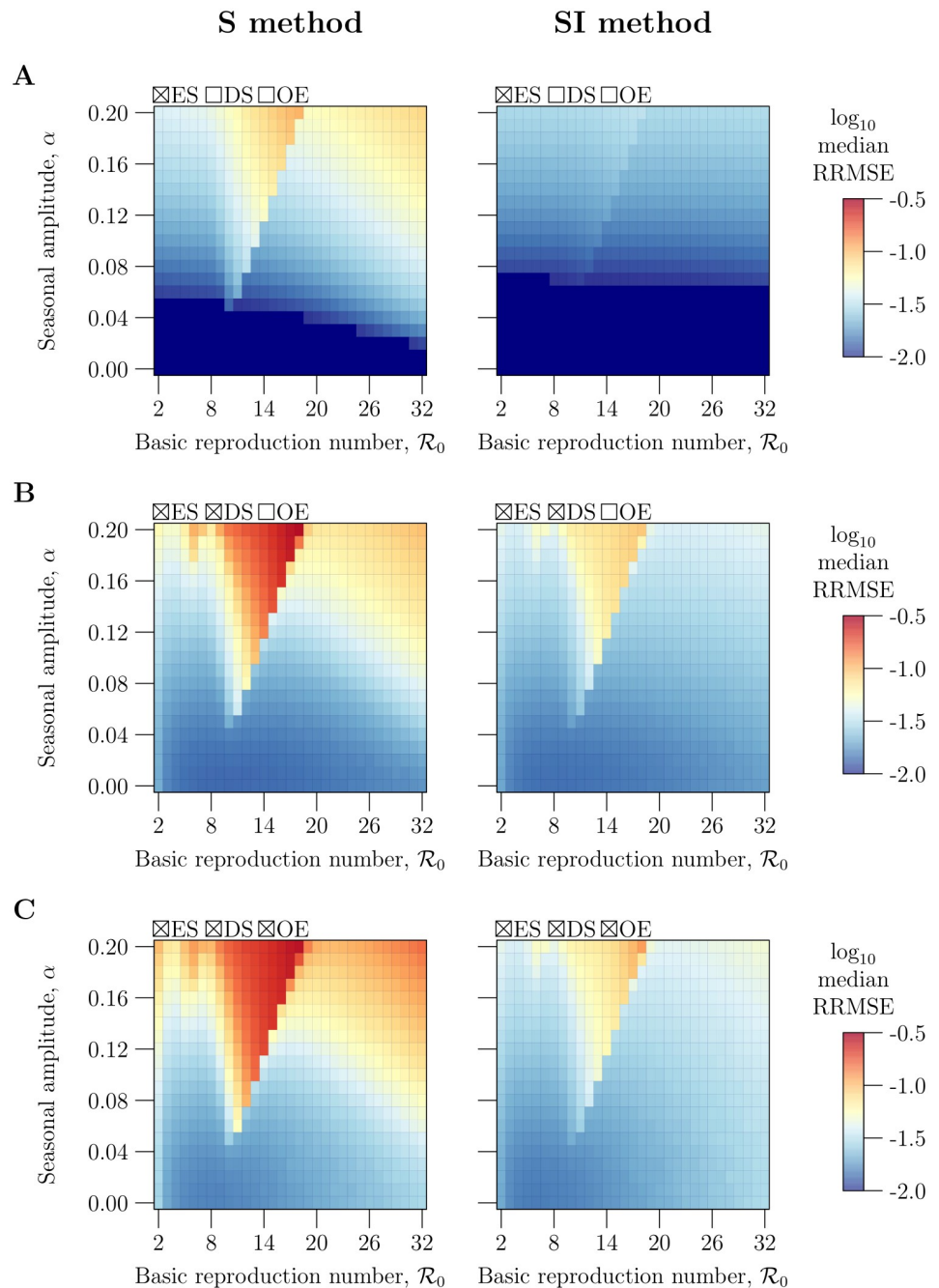
Fig 6 plots median RRMSE as a univariate function of each of 6 additional parameters—the initial states  $S_0$  and  $I_0$ , vital rates  $\nu_c$  and  $\mu_c$ , mean generation interval  $t_{\text{gen}}$ , and case reporting probability  $p_{\text{rep}}$ —with the focal parameter assigned values between  $\frac{1}{4}$  and 4 times its reference value (Table 1). The horizontal axis measures the ratio of the focal parameter's data-generating value to its reference value, so that commensurate deviations from the reference case can be compared across the 6 parameters.

In order to produce Figs 5 and 6, we assigned reference values (Table 1) to all but the focal data-generating parameter(s) (e.g., all except  $\langle\beta\rangle$  and  $\alpha$  in Fig 5). We fit loess curves  $\beta_{\text{loess}}(t; q)$  to all raw estimates  $\beta_k$  of  $\beta(t)$ , and recorded the RRMSE in  $\beta_{\text{loess}}(t_k; q)$ . Motivated by Fig 4B and Eq (59), we fixed  $q = q^*$ , taking  $q^* = 65$  with the S method and  $q^* = 53$  with the SI method.

A pattern in our interpretation of Figs 5 and 6 below is that error in  $\beta(t)$  estimation is sensitive to a parameter if changes in that parameter (i) cause incidence  $Z(t)$  or prevalence  $I(t)$  to approach zero more frequently or more closely, or (ii) increase noise in estimated incidence  $Z_k$  or estimated prevalence  $I_k$ . Both outcomes incorrectly increase noise in  $\beta_k$  (cf. §3.2).

When the noise in  $\beta_k$  is extreme, setting  $q = q^*$  can undersmooth the time series ( $q^* < q_{\text{opt}}$ ). In this case, smaller RRMSE is attainable by determining  $q_{\text{opt}}$  and setting  $q = q_{\text{opt}}$ . Nevertheless, we did not find  $q_{\text{opt}}$  for each of the  $5 \times 10^6$  time series considered by Figs 5 and 6, which would have increased the total computation time by a factor of 100. Consequently, Figs 5 and 6 may overestimate the sensitivity of  $\beta(t)$  estimation error to data-generating parameters. (In §S5.3 of S1 Text, we show that the quantitative effect of choosing  $q^*$  over  $q_{\text{opt}}$  is likely to be small.)

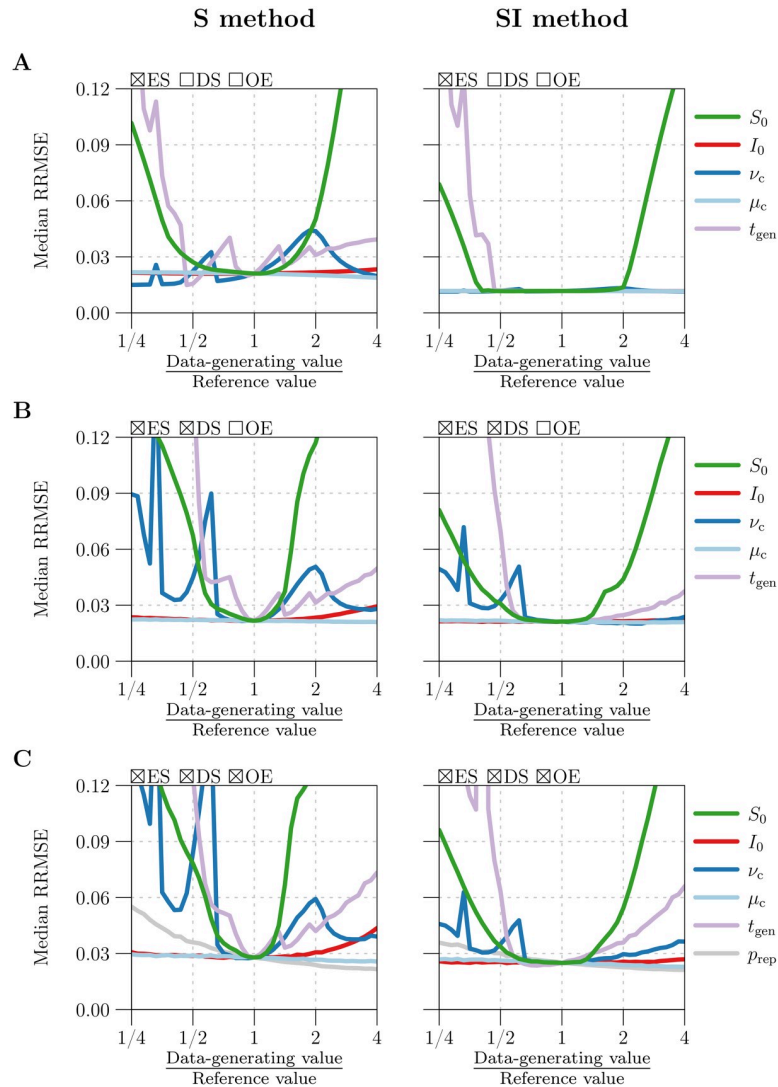
**3.5.1 Sensitivity to the basic reproduction number  $\mathcal{R}_0$  and seasonal amplitude  $\alpha$  (Fig 5).** For fixed  $\alpha$ , median RRMSE was a non-monotonic function of  $\mathcal{R}_0$ . The reason is that



**Fig 5. Sensitivity of  $\beta(t)$  estimation error to the mean  $\langle\beta\rangle$  and amplitude  $\alpha$  of seasonal forcing.** Contained in each panel are heatmaps of median RRMSE (Eq (33)) in estimates of a seasonally forced  $\beta(t)$  (Eq (27)) from simulated reported incidence time series, as a bivariate function of the mean  $\langle\beta\rangle$  and amplitude  $\alpha$  of seasonal forcing. The  $\langle\beta\rangle$  axis has been scaled to measure the basic reproduction number  $\mathcal{R}_0$  (Eq (2)). When simulating reported incidence, reference values (Table 1) were assigned to all data-generating parameters except  $\langle\beta\rangle$  and  $\alpha$ . A grid of  $(\mathcal{R}_0, \alpha)$  pairs with levels  $\mathcal{R}_0 = 2, 3, \dots, 32$  and  $\alpha = 0, 0.01, \dots, 0.2$  was considered, with  $\langle\beta\rangle$  defined for each value of  $\mathcal{R}_0$  via Eq (2). For each parametrization, 1000 simulations were performed with environmental stochasticity [ES] ( $\epsilon = 0.5$ ) and with or without demographic stochasticity [DS] and observation error [OE], as indicated by row: [Row A] without DS or OE ( $p_{\text{rep}} = 1, t_{\text{rep}} = 0$  weeks), [Row B] with DS but without OE ( $p_{\text{rep}} = 1, t_{\text{rep}} = 0$  weeks), [Row C] with DS and OE ( $p_{\text{rep}} = 0.25, t_{\text{rep}} = 2$  weeks). Corresponding mock birth and natural mortality time series were created, then  $\beta(t)$  was estimated from the data using [Left] the S method and [Right] the SI method, all without input error. For each set of estimates of  $\beta(t)$  (1000 estimates per parametrization, per simulation method, per estimation method), the median RRMSE was calculated (after smoothing with fixed  $q$ ; see Eq (59)) and displayed as one point in the appropriate heatmap, coloured according to the logarithmic scale on the right. The darkest blue indicates median RRMSE less than 0.01.

<https://doi.org/10.1371/journal.pcbi.1008124.g005>





**Fig 6. Sensitivity of  $\beta(t)$  estimation error to data-generating parameters other than  $\beta$  and  $\alpha$ .** Plotted in each panel is the median RRMSE (Eq (33)) in estimates of a seasonally forced  $\beta(t)$  (Eq (27)) from simulated reported incidence time series ( $\Delta t = 1$  week,  $n = 1042$ ), as a univariate function of each of 5 or 6 data-generating parameters (indicated by the legend). When simulating reported incidence, reference values (Table 1) were assigned to all but the focal parameter, which was assigned 41 values logarithmically spaced between  $\frac{1}{4}$  and 4 times its reference value. The horizontal axis (logarithmic scale) measures the ratio of the focal parameter's true value to its reference value, so that commensurate deviations from the reference case can be compared across parameters. For each parametrization, 1000 simulations were performed with environmental stochasticity [ES] ( $\epsilon = 0.5$ ) and with or without demographic stochasticity [DS] and observation error [OE], as indicated by row: **[Row A]** without DS or OE ( $p_{\text{rep}} = 1$ ,  $t_{\text{rep}} = 0$  weeks), **[Row B]** with DS but without OE ( $p_{\text{rep}} = 1$ ,  $t_{\text{rep}} = 0$  weeks), or **[Row C]** with DS and OE ( $p_{\text{rep}} = 0.25$  except when  $p_{\text{rep}}$  is the focal parameter,  $t_{\text{rep}} = 2$  weeks). Corresponding mock birth and natural mortality time series were created, then  $\beta(t)$  was estimated from the data using **[Left]** the S method and **[Right]** the SI method, all without input error. For each set of estimates of  $\beta(t)$  (1000 estimates per parametrization, per simulation method, per estimation method), the median RRMSE was calculated (after smoothing with fixed  $q$ ; see Eq (59)) and displayed as one point in the appropriate panel and graph.

changes in (effective)  $\mathcal{R}_0$  are responsible for dynamical transitions that alter the structure of solutions of the SIR model (1) [28, 42, 43]. Specifically, as  $\mathcal{R}_0$  is increased from 2 to 32, minimum incidence  $Z_{\text{min}}$  and minimum prevalence  $I_{\text{min}}$  on the attractor varies non-monotonically (see Fig 2 in [28]). Smaller  $Z_{\text{min}}$  and  $I_{\text{min}}$  yield more noise in  $\beta_k$  and correspondingly greater

RRMSE. For fixed  $\mathcal{R}_0$ ,  $I_{\min}$  decreases monotonically as  $\alpha$  is increased from 0 to 1 (see Fig 11 in [43]), so we expect median RRMSE to increase monotonically with  $\alpha$ , as observed in Fig 5.

**3.5.2 Sensitivity to the initial state ( $S_0, I_0$ ) (Fig 6).** RRMSE is sensitive to the data-generating  $S_0$ , but not  $I_0$ . The reference values of  $S_0$  and  $I_0$  are taken from a point  $(S^*, I^*, R^*)$  on the attractor of the SIR model (1) with seasonally forced  $\beta(t)$  and constant vital rates  $\nu_c$  and  $\mu_c$  (cf. §2.3.4). When  $S_0$  is far from  $S^*$ , the solution of system (1) undergoes extreme fluctuation before relaxing to the attractor, and both  $Z$  and  $I$  approach zero during the transient, generating spurious noise at the start of the  $\beta_k$  time series.

Note that  $I_0$  differing from  $I^*$  has a much smaller effect on dynamics than  $S_0$  differing from  $S^*$  by the same factor. Since  $I^* \ll S^*$ , the perturbation of  $(S_0, I_0, R_0)$  from the attractor is much smaller.

**3.5.3 Sensitivity to vital rates  $\nu_c$  and  $\mu_c$  (Fig 6).** Median RRMSE was a non-monotonic function of the data-generating birth rate  $\nu_c$ . This behaviour arises because scaling  $\nu_c$  is dynamically equivalent to scaling  $\mathcal{R}_0$  by the same factor [2, 28], and median RRMSE is a non-monotonic function of  $\mathcal{R}_0$  (cf. §3.5.1 above).

Changing the data-generating natural mortality rate  $\mu_c$  had a negligible effect on RRMSE. This is unsurprising, because natural death is dominated by recovery and disease-induced death in governing the rate of infected decrease. That is,  $\gamma \gg \mu(t)$  in Eq (1b), so changes in  $\mu_c$  by up to a factor of 4 have little effect on dynamics.

**3.5.4 Sensitivity to the mean generation interval  $t_{\text{gen}}$  (Fig 6).** Median RRMSE increased rapidly as the data-generating  $t_{\text{gen}}$  was made smaller than  $2^{-4/5}$  (roughly 0.57) times its reference value of 13 days. A period-doubling bifurcation occurs near this value of  $t_{\text{gen}}$ , and the attractor of the SIR model (1) acquires a 2-year cycle with much smaller  $Z_{\min}$  and  $I_{\min}$  (see §5.3.1 of S1 Text). Propagation of noise to  $\beta_k$  intensifies, resulting in greater RRMSE.

The performance of the S method fluctuates more as a function of  $t_{\text{gen}}$  than that of the SI method. This occurs because the S method rounds  $t_{\text{gen}}$  in the numerator of Eq (25c) to the nearest integer multiple of  $\Delta t$ , and the rounding error oscillates as a function of  $t_{\text{gen}}$ . The SI method does not require rounding, so these fluctuations are not observed.

**3.5.5 Sensitivity to the case reporting probability  $p_{\text{rep}}$  (Fig 6).** When the reported incidence data contain observation error (Fig 6C), RRMSE is additionally sensitive to the case reporting probability  $p_{\text{rep}}$ . Decreasing  $p_{\text{rep}}$  increases noise in reported incidence  $C_k$  (Eq (31)), which is propagated to estimated incidence  $Z_k$ , estimated prevalence  $I_k$ , and in turn  $\beta_k$  (cf. §3.2).

Fig 6 suggests weak sensitivity to  $p_{\text{rep}}$ . However, noise in  $Z_k$  and  $I_k$  is amplified in  $\beta_k$  to the extent that  $Z$  and  $I$  are close to zero (cf. §3.2). Hence, for example, if the data-generating  $t_{\text{gen}}$  were assigned a value smaller than half its reference value of 13 days, then we would have observed more acute sensitivity to  $p_{\text{rep}}$  as a result of closer approaches to zero by  $Z$  and  $I$  (cf. §3.5.4 above).

**3.5.6 S method versus SI method (Figs 5 and 6).** Both the S and SI methods performed well, estimating  $\beta(t)$  with median RRMSE less than 10% across most parametrizations. However, by resisting noise propagation (cf. §3.2), the SI method was significantly less sensitive to the data-generating parameters and to the addition of demographic stochasticity and observation error.

### 3.6 Sensitivity to mis-specification of input parameters

In §3.5, we considered the ideal situation in which the user knows the true (data-generating) values of the input parameters. Here, we examine the more realistic situation in which the

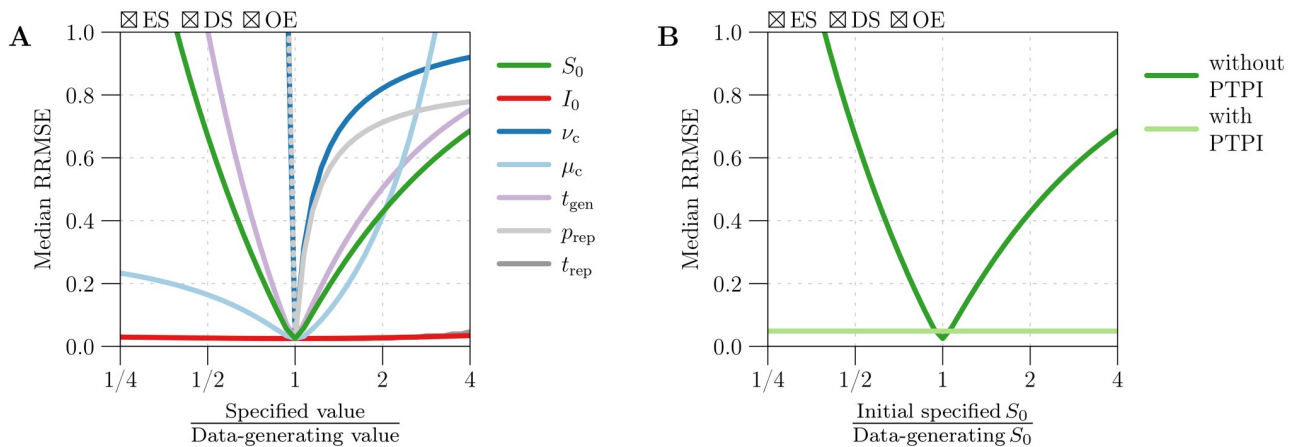
user specifies input parameters with some error. The effect of mis-specification is particularly important for parameters that are difficult to estimate accurately, such as the case reporting probability  $p_{\text{rep}}$ . The details of our analysis are outlined in §2.6.2.

We restrict our attention to application of the SI method to reported incidence data simulated with process and observation error. Differences in RRMSE between methods of data simulation and  $\beta(t)$  estimation are dominated (by an order of magnitude) by the increase in RRMSE resulting from mis-specified input parameters.

Fig 7A plots the median RRMSE in estimates of  $\beta(t)$  from 1000 realizations of a reported incidence time series, as a univariate function of the factor by which an input parameter—one of the initial states  $S_0$  and  $I_0$ , mean generation interval  $t_{\text{gen}}$ , vital rates  $\nu_c$  and  $\mu_c$ , and case reporting parameters  $p_{\text{rep}}$  and  $t_{\text{rep}}$ —was mis-specified. The specified value of the focal parameter was varied between  $\frac{1}{4}$  and 4 times its true (data-generating) value, and the remaining parameters were specified without error.

**3.6.1 Sensitivity to error in the specified initial state ( $S_0, I_0$ ).** Fig 7 shows that error in the specified value of  $S_0$  is propagated non-negligibly to estimates of  $\beta(t)$ , while mis-specification of  $I_0$  has practically no effect on  $\beta(t)$  estimation error. Eqs (40) and (41) show that specifying incorrect values  $S'_0$  and  $I'_0$  for  $S_0$  and  $I_0$  creates errors in  $S_k$  and  $I_k$  that vanish geometrically as  $k \rightarrow \infty$ . However, since  $t_{\text{life}} \gg t_{\text{inf}}$ , the decay is significantly slower in  $S_k$ . Indeed, with reference values  $\mu_c = 0.04 \text{ year}^{-1}$ ,  $t_{\text{gen}} = \gamma^{-1} = 13 \text{ days}$ , and  $\Delta t = 1 \text{ week}$ , we find that a factor of 10 reduction in error between times  $t_k$  and  $t_{k+i}$  requires just  $i = 5$  in the infected time series, compared to  $i = 3002$  in the susceptible time series (roughly 58 years with  $\Delta t = 1 \text{ week}$ ). Hence, in practice, accurate reconstruction of  $S(t)$ ,  $I(t)$ , and in turn  $\beta(t)$  relies on accurate specification of  $S_0$ , but not  $I_0$ . We address sensitivity to mis-specification of  $S_0$  in §3.7 below.

**3.6.2 Sensitivity to error in the specified birth rate  $\nu_c$  and case reporting probability  $p'_{\text{rep}}$**  Mis-specifying  $\nu_c$  or  $p_{\text{rep}}$  by a factor of  $2^{1/10}$  (7.2%) yielded median RRMSE greater than 30%. Mis-specifying by a factor of  $2^{-1/10}$  (−6.7%) led to even worse estimates of  $\beta(t)$ , with median



**Fig 7. Sensitivity of  $\beta(t)$  estimation error to the user-specified values of input parameters.** [Panel A] Median RRMSE (Eq (33)) in estimates of  $\beta(t)$  from simulated reported incidence time series ( $\Delta t = 1 \text{ week}$ ,  $n = 1042$ ), as a univariate function of the factor by which an input parameter was mis-specified. One thousand simulations were performed using fixed values (Table 1) for all data-generating parameters. The simulations accounted for environmental stochasticity [ES] ( $\epsilon = 0.5$ ), demographic stochasticity [DS], and observation error [OE] ( $p_{\text{rep}} = 0.25$ ,  $t_{\text{rep}} = 2 \text{ weeks}$ ). For each simulation, corresponding mock birth and natural mortality time series were created, and  $\beta(t)$  was estimated from the data using the SI method. True (data-generating) values were specified for all input parameters except the focal parameter (indicated by the legend), for which 41 values logarithmically spaced between  $\frac{1}{4}$  and 4 times the true value were specified in turn. Each input parametrization yielded 1000 estimates of  $\beta(t)$ , whose median RRMSE was calculated (after smoothing with fixed  $q$ ; see Eq (59)) and displayed as one point in the appropriate graph. [Panel B] Result of repeating the analysis from Panel A in which  $S_0$  was specified with varying amounts of error, but with the initially erroneous value of  $S_0$  updated using the method of peak-to-peak iteration (PTPI; 25 iterations) prior to  $\beta(t)$  estimation. The original result, obtained without PTPI, is presented for comparison.

<https://doi.org/10.1371/journal.pcbi.1008124.g007>

RRMSE exceeding 100% (not visible in Fig 7A). Eqs (42) and (47) show that specifying incorrect values  $v'_c$  and  $p'_{\text{rep}}$  for  $v_c$  and  $p_{\text{rep}}$  generates absolute errors in  $S_k$  that tend to increase over time ( $k$ ) to a limit. In practice, systematic underestimation of births by the  $B_k$  time series (modeled here by  $v'_c < v_c$ ) and overestimation of incidence by the  $Z_k$  time series ( $p'_{\text{rep}} < p_{\text{rep}}$ ) can cause  $S_k$  to eventually take negative values. Once this happens, attempts by the S and SI methods to reconstruct  $\beta(t)$  fail completely.

While this failure may seem concerning, it should be viewed as a tool for diagnosing incorrect birth and case reporting rates: if the S or SI method yields negative  $S_k$  for any  $k$ , then one should speculate that births were underestimated or that incidence was overestimated, and retry the algorithm with a scaled up  $B_k$  time series and/or with greater  $p_{\text{rep}}$  (as  $Z_k$  is computed by scaling reported incidence by a factor of  $p_{\text{rep}}^{-1}$ ; see Eqs (25a) and (26a)). Of course, overcorrection is also undesirable (*cf.* right half of Fig 7A). In our work, we have found that a brief exploration of possible adjustments—factors by which to increase  $B_k$  and/or  $p_{\text{rep}}$ —suffices to identify ones that prevent both negative  $S_k$  and pronounced transient dynamics at the start of the susceptible time series (indicating under- or overcorrection).

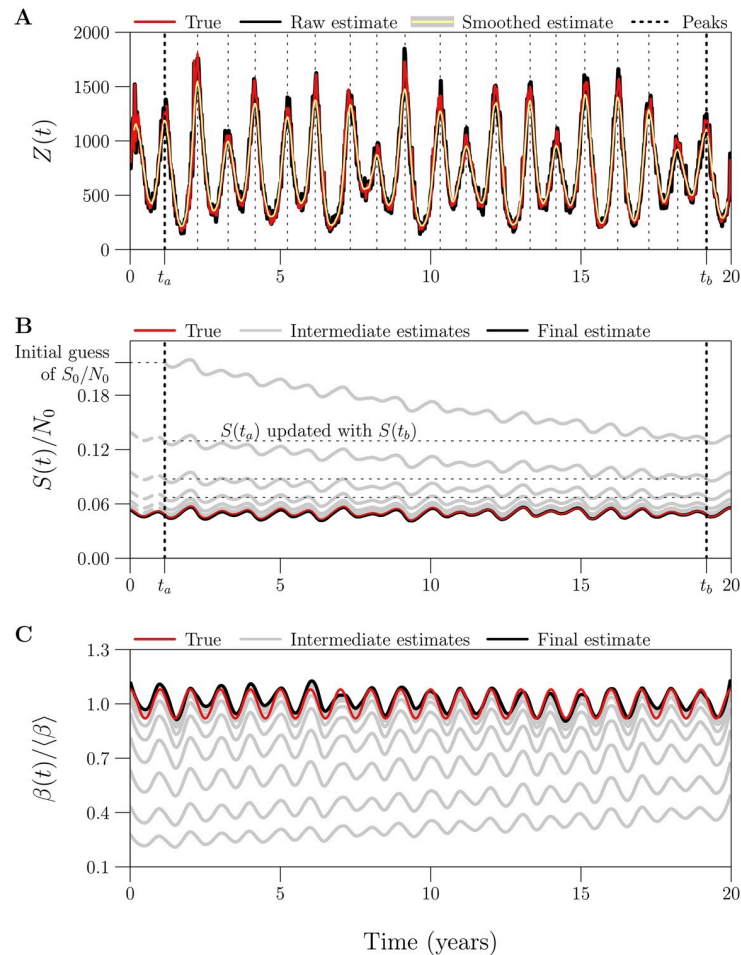
### 3.7 Solution of the $S_0$ estimation problem using PTPI

In §3.6, we showed that the performance of the S and SI methods is highly sensitive to misspecification of the initial number of susceptibles  $S_0$ . Here, we assess PTPI as a way to iteratively improve initially poor estimates of  $S_0$  prior to reconstruction of  $S(t)$  and  $\beta(t)$ .

Fig 8 demonstrates PTPI for an example in which  $S_0$  was overestimated by a factor of 4 by a user of the SI method. PTPI yielded increasingly accurate estimates of  $S_0$  and correspondingly more accurate reconstructions of  $S(t)$  (Fig 8B) and  $\beta(t)$  (Fig 8C). Fig 7B repeats our analysis from §3.6, except using PTPI (25 iterations) to update the incorrect estimate of  $S_0$  prior to reconstructing  $\beta(t)$ . We see that application of PTPI in conjunction with the SI method enables accurate  $\beta(t)$  reconstruction independently of errors in the initial estimate of  $S_0$ . This result is unsurprising in light of Fig 9, which shows that PTPI converges rapidly (in fewer than 10 iterations) to an accurate estimate of  $S_0$  independently of the initial guess. Due to process error in the underlying dynamics, the relative error in the limiting estimate of  $S_0$  varied between the 1000 realizations of reported incidence considered (5th–95th percentile range [−11.9, 12.5]%, median 0.9%). Process error creates variance in the time between peaks in incidence (see Fig 8A), violating the periodicity assumption of PTPI (the theoretical basis of the technique; *cf.* §2.8). Nevertheless, Figs 7–9 demonstrate that PTPI can significantly improve  $S(t)$  and  $\beta(t)$  reconstruction from roughly periodic incidence data.

### 3.8 Run time

We implemented the S and SI methods and PTPI in R and ran them on a MacBook Pro with a 2.4 GHz Quad-Core Intel Core i5 chip. The S and SI methods are both extremely fast, requiring a total of 0.124 and 0.376 seconds, respectively, to generate a reconstruction of  $\beta(t)$  from 1000 years of weekly reported incidence ( $\Delta t = 1$  week,  $n = 52142$ ). Application of PTPI in conjunction with either method increases the run time with each iteration, but the total run time remains inconsequential due to the rate of convergence of the iterations to a limiting estimate of  $S_0$ . For example, when we applied PTPI to the same simulated data, the truncation step (Box 5) added 0.094 seconds to the total run time, while the iteration step (Box 5) added 1.01 seconds per iteration on average.



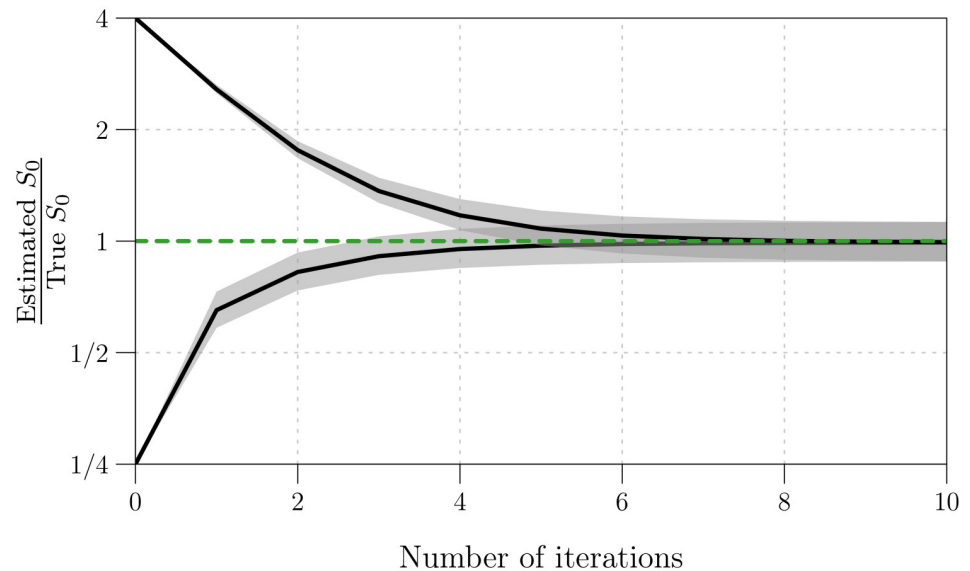
**Fig 8. Example of  $S(t)$  and  $\beta(t)$  reconstruction with an overestimate of  $S_0$  corrected by peak-to-peak iteration (PTPI).** [Panel A] Truncation step of PTPI (Box 5). Plotted is a reconstruction of true incidence  $Z(t)$  from a simulated reported incidence time series, before [ $Z_k$ , black] and after [ $\bar{Z}_k$ , yellow] smoothing with a 13-point central moving average. Vertical lines indicate peaks in  $\bar{Z}_k$ . The times of the first peak in  $\bar{Z}_k$  and the last peak occurring at the same phase of the cycle (in this case, the last peak) are denoted by  $t_a$  and  $t_b$ . [Panel B] Iteration step of PTPI (Box 6), where the initial estimates of both  $S_0 = S(0)$  and  $S(t_a)$  were taken to be 4 times the true (data-generating) value of  $S_0$ . Plotted in grey are successive reconstructions of  $S(t)$  between times  $t_a$  and  $t_b$ , generated by updating the estimate of  $S(t_a)$  with the estimate of  $S(t_b)$  obtained in the previous iteration. Dashed continuations to the left of  $t_a$  display estimation of  $S_0$  backwards in time from estimates of  $S(t_a)$ . Plotted in black is the result of reconstructing  $S(t)$  starting from the final estimate of  $S_0$ , which was obtained after 25 iterations and had a relative error of roughly 1.4% (compared to 300% in the initial estimate). [Panel C] The sequence of reconstructions of  $\beta(t)$  corresponding to the estimates of  $S_0$  shown in Panel B. [Details] Twenty years of weekly reported incidence ( $\Delta t = 1$  week,  $n = 1042$ ) were simulated with environmental noise in transmission ( $\epsilon = 0.5$ ), demographic stochasticity, and random under-reporting of cases ( $p_{\text{rep}} = 0.25$ ), using reference values (Table 1) for the remaining parameters.  $Z(t)$ ,  $S(t)$  and  $\beta(t)$  were reconstructed from reported incidence using the SI method without input error (apart from mis-specification of  $S_0$ ).

<https://doi.org/10.1371/journal.pcbi.1008124.g008>

## 4 Discussion

We have compared three fast methods of estimating the time-varying transmission rate  $\beta(t)$  from reported incidence time series, all based on discretizations of the SIR model (1). Fine and Clarkson's method [6], referred to here as the FC method, fails rapidly in practice, because it treats natural mortality in the susceptible population as negligible. Although Krylova's method [24], adapted here as the S method, corrects this limitation of the FC method and is accurate





**Fig 9. Convergence of estimates of  $S_0$  obtained using peak-to-peak iteration (PTPI).**  $S_0$  was estimated by applying PTPI (25 iterations) to 1000 incidence time series (*i.e.*, 1000 realizations of a reported incidence time series, scaled by  $p_{\text{rep}}^{-1}$ ). An initial guess for  $S_0$  was taken to be  $\frac{1}{4}$  or 4 times the true (data-generating) value. For each initial guess, this process generated 1000 sequences of 26 estimates of  $S_0$ . Plotted are the median [black lines] and 5th–95th percentile range [grey bands] of the estimate of  $S_0$  at each iteration, for the first 10 iterations. The vertical axis measures (on a logarithmic scale) the ratio of the estimated and true values of  $S_0$ , hence convergence close to 1 [dashed green line] represents convergence of the estimates close to the true value. [Details] One thousand reported incidence time series ( $\Delta t = 1$  week,  $n = 1042$ ) were simulated with environmental noise in transmission ( $\epsilon = 0.5$ ), demographic stochasticity, and random under-reporting of cases ( $p_{\text{rep}} = 0.25$ ), using reference values (Table 1) for the remaining parameters, including  $S_0$  (hence  $S_0$  was the same in all simulations). True incidence was estimated from reported incidence via Eq (26a) (with reporting parameters  $p_{\text{rep}}$  and  $t_{\text{rep}}$  correctly specified), yielding 1000 time series of estimated incidence. Corresponding mock (constant) birth and natural mortality time series were created (with vital rates  $\nu_c$  and  $\mu_c$  correctly specified), and these data (estimated incidence, births, natural mortality) were passed to the PTPI algorithm, allowing for iterative re-estimation of  $S_0$ .

<https://doi.org/10.1371/journal.pcbi.1008124.g009>

for certain simulated data, her method suffers from extreme sensitivity to process and observation error. Specifically, noise in reported incidence is spuriously propagated to its estimates of  $\beta(t)$ . Our algorithm for transmission rate estimation, referred to here as the SI method and based on deJonge’s method [25], is much more resilient to noise in reported incidence and therefore superior to the S method.

Like its predecessors, the SI method is sensitive to (i) certain input parameters: the initial number of susceptible individuals  $S_0$ , the case reporting probability  $p_{\text{rep}}$ , and the mean generation interval  $t_{\text{gen}}$ ; as well as (ii) vital data: times series of births and natural mortality without substantial systematic errors.

The requirement of a good estimate of  $S_0$  has been a major barrier to use of existing fast methods of  $\beta(t)$  estimation (including those presented in [6, 24, 25]). We have proposed and demonstrated PTPI as a valid and fast technique for obtaining accurate estimates of  $S_0$  from poor initial guesses, conditional on periodic dynamics (epidemic recurrence with a fixed period). Use of the SI method in conjunction with PTPI represents a major advance over the existing fast methods.

Estimation of the case reporting probability  $p_{\text{rep}}$  is possible using maximum likelihood approaches, including trajectory matching. However, a fast way to obtain a crude estimate of  $p_{\text{rep}}$  is to divide cumulative reported incidence over the time interval  $[t_0, t_n]$ , by the cumulative

incidence that is expected from the unforced SIR model (system (1) with  $\beta \equiv \langle \beta \rangle$ ,  $\nu \equiv \nu_c$ , and  $\mu \equiv \mu_c$ ) at equilibrium:

$$p_{\text{rep}} \approx \frac{\sum_{k=1}^n C_k}{\nu_c \tilde{N}_0 (1 - \frac{1}{R_0})(t_n - t_0)}. \quad (60)$$

This approximation can be made in temporal subintervals to obtain a time-varying reporting rate, which would replace the constant  $p_{\text{rep}}$  in Eq (26a). Sensitivity of the SI method to misspecification of the mean generation interval ( $t_{\text{gen}}$ ) may be of greater concern, though if the distribution of the incubation period (time from infection to onset of symptoms) is narrow, then  $t_{\text{gen}}$  will be well approximated by the (observable) mean serial interval [44].

Overall, the SI method, in conjunction with PTPI, represents a highly tractable approach to reconstructing susceptibles and  $\beta(t)$  from infectious disease time series that span decades or centuries. It makes fewer assumptions about the disease and population of interest than the regression-based tSIR method [7, 23] (*i.e.*, it does not require an infectious period equal to the observation interval, ignore susceptible mortality, or assume that cumulative incidence approximates cumulative births). Moreover, it is significantly less complex and much less computationally demanding than simulation-based methods of inference, such as iterated filtering [8, 19, 20] and generalized profiling [21, 22].

Even when the observed infectious disease time series is short enough that simulation-based methods are tractable, the approach to transmission rate reconstruction that we promote here can be usefully employed to provide better starting conditions at negligible computational cost.

## Supporting information

**S1 Text. Text supplement.** A .pdf document containing annotated R code, making the results reported here completely reproducible by the reader.  
(PDF)

**S1 File. Source files.** A .zip archive containing all of the source files needed to compile S1 Text.  
(ZIP)

## Acknowledgments

We thank Ben Bolker, Jonathan Dushoff, and Sang Woo Park for helpful comments and discussion.

## Author Contributions

**Conceptualization:** Mikael Jagan, Michelle S. deJonge, Olga Krylova, David J. D. Earn.

**Formal analysis:** Mikael Jagan, Michelle S. deJonge, David J. D. Earn.

**Funding acquisition:** Mikael Jagan, Olga Krylova, David J. D. Earn.

**Investigation:** Mikael Jagan, Michelle S. deJonge, David J. D. Earn.

**Methodology:** Mikael Jagan, Michelle S. deJonge, Olga Krylova, David J. D. Earn.

**Project administration:** Mikael Jagan, Michelle S. deJonge, David J. D. Earn.

**Resources:** David J. D. Earn.

**Software:** Mikael Jagan.

**Supervision:** David J. D. Earn.

**Validation:** Mikael Jagan.

**Visualization:** Mikael Jagan.

**Writing – original draft:** Mikael Jagan, Michelle S. deJonge.

**Writing – review & editing:** Mikael Jagan, David J. D. Earn.

## References

1. Dietz K. The incidence of infectious diseases under the influence of seasonal fluctuations. In: *Mathematical Models in Medicine*. vol. 11 of *Lecture Notes in Biomathematics*. Springer-Verlag Berlin / Hiedelberg; 1976. p. 1–15.
2. Earn DJD, Rohani P, Bolker BM, Grenfell BT. A simple model for complex dynamical transitions in epidemics. *Science*. 2000; 287(5453):667–670. <https://doi.org/10.1126/science.287.5453.667> PMID: 10650003
3. Shaman J, Kohn M. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*. 2009; 106(9):3243–3248. <https://doi.org/10.1073/pnas.0806852106>
4. London W, Yorke JA. Recurrent outbreaks of measles, chickenpox and mumps. I. Seasonal variation in contact rates. *American Journal of Epidemiology*. 1973; 98(6):453–468. <https://doi.org/10.1093/oxfordjournals.aje.a121575> PMID: 4767622
5. Hethcote HW. The mathematics of infectious diseases. *SIAM Review*. 2000; 42(4):599–653. <https://doi.org/10.1137/S0036144500371907>
6. Fine PEM, Clarkson JA. Measles in England and Wales—I: an analysis of factors underlying seasonal patterns. *International Journal of Epidemiology*. 1982; 11(1):5–14. <https://doi.org/10.1093/ije/11.1.5> PMID: 7085179
7. Finkenstädt B, Grenfell B. Time series modelling of childhood diseases: a dynamical systems approach. *Journal of the Royal Statistical Society C (Applied Statistics)*. 2000; 49(2):187–205. <https://doi.org/10.1111/1467-9876.00187>
8. He D, Ionides EL, King AA. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface*. 2010; 7:271–283. <https://doi.org/10.1098/rsif.2009.0151>
9. Hempel K, Earn DJD. A century of transitions in New York City's measles dynamics. *Journal of the Royal Society Interface*. 2015; 12(106):20150024. <https://doi.org/10.1098/rsif.2015.0024>
10. Pollicott M, Wang H, Weiss H. Extracting the time-dependent transmission rate from infection data via solution of an inverse ODE problem. *Journal of Biological Dynamics*. 2012; 6(2):509–523. <https://doi.org/10.1080/17513758.2011.645510> PMID: 22873603
11. Lange A. Reconstruction of disease transmission rates: applications to measles, dengue, and influenza. *Journal of Theoretical Biology*. 2016; 400:138–153. <https://doi.org/10.1016/j.jtbi.2016.04.017> PMID: 27105674
12. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*. 2004; 160:509–516. <https://doi.org/10.1093/aje/kwh255> PMID: 15353409
13. Smirnova A, deCamp L, Chowell G. Forecasting epidemics through nonparametric estimation of time-dependent transmission rates using the SEIR model. *Bulletin of Mathematical Biology*. 2019; 81:4343–4365. <https://doi.org/10.1007/s11538-017-0284-3> PMID: 28466232
14. Tien JH, Poinar HN, Fisman DN, Earn DJD. Herald waves of cholera in nineteenth century London. *Journal of the Royal Society Interface*. 2011; 8(58):756–760. <https://doi.org/10.1098/rsif.2010.0494>
15. Krylova O, Earn DJD. Patterns of smallpox mortality in London, England, over three centuries. *PLoS Biology*. 2020 <https://doi.org/10.1371/journal.pbio.3000506>
16. Anderson RM, May RM. *Infectious Diseases of Humans: Dynamics and Control*. Oxford, UK: Oxford University Press; 1991.



17. Morton A, Finkenstädt B. Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society C (Applied Statistics)*. 2005; 54(3):575–594. <https://doi.org/10.1111/j.1467-9876.2005.05366.x>
18. Cauchemez S, Ferguson NM. Likelihood-based estimation of continuous-time epidemic models from time series data: application to measles transmission in London. *Journal of the Royal Society Interface*. 2008; 5(25):885–897. <https://doi.org/10.1098/rsif.2007.1292>
19. Ionides EL, Breto C, King AA. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*. 2006; 103(49):18438–18443. <https://doi.org/10.1073/pnas.0603181103>
20. King AA, Nguyen D, Ionides EL. Statistical inference for partially observed Markov processes via the R package pomp. *Journal of Statistical Software*. 2009; 69(12):1–43.
21. Ramsay JO, Hooker G, Campbell D, Cao J. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society B (Statistical Methodology)*. 2007; 69(5):741–796. <https://doi.org/10.1111/j.1467-9868.2007.00610.x>
22. Hooker G, Ellner SP, De Vargas Roditi L, Earn DJD. Parameterizing state-space models for infectious disease dynamics by generalized profiling: measles in Ontario. *Journal of the Royal Society Interface*. 2011; 8(60):961–974. <https://doi.org/10.1098/rsif.2010.0412>
23. Becker A, T GB. tsiR: An R package for time series susceptible-infected-recovered models of epidemics. *PLoS ONE*. 2017; 12(9):0185528. <https://doi.org/10.1371/journal.pone.0185528>
24. Krylova O. Predicting epidemiological transitions in infectious disease dynamics. Smallpox in historic London (1664–1930). Hamilton, Ontario, Canada: McMaster University; 2011. Available from: <https://macsphere.mcmaster.ca/handle/11375/11231>.
25. deJonge MS. Fast estimation of time-varying transmission rates. Hamilton, Ontario, Canada: McMaster University; 2014. Available from: <https://macsphere.mcmaster.ca/handle/11375/14230>.
26. 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>
27. Champredon D, Dushoff J. Intrinsic and realized generation intervals in infectious-disease transmission. *Proceedings of the Royal Society B (Biological Sciences)*. 2015; 282(1821):20152026. <https://doi.org/10.1098/rspb.2015.2026>
28. Krylova O, Earn DJD. Effects of the infectious period distribution on predicted transitions in childhood disease dynamics. *Journal of the Royal Society Interface*. 2013; 10:20130098. <https://doi.org/10.1098/rsif.2013.0098>
29. Brauer F, Castillo-Chavez C. *Mathematical models in population biology and epidemiology*. New York, NY: Springer; 2012.
30. Lloyd AL. Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. *Proceedings of the Royal Society B (Biological Sciences)*. 2001; 268(1470):985–993. <https://doi.org/10.1098/rspb.2001.1599>
31. Lloyd AL. Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. *Theoretical Population Biology*. 2001; 60(1):59–71. <https://doi.org/10.1006/tpbi.2001.1525> PMID: 11589638
32. Ma J, Ma Z. Epidemic threshold conditions for seasonally forced SEIR models. *Mathematical Biosciences and Engineering*. 2006; 3(1):161–172. <https://doi.org/10.3934/mbe.2006.3.161> PMID: 20361816
33. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons; 2019.
34. Goldstein E, Dushoff J, Ma J, Plotkin JB, Earn DJD, Lipsitch M. Reconstructing influenza incidence by deconvolution of daily mortality time series. *Proceedings of the National Academy of Sciences*. 2009; 106:21825–21829. <https://doi.org/10.1073/pnas.0902958106>
35. He D, Earn DJD. The cohort effect in childhood disease dynamics. *Journal of the Royal Society Interface*. 2016; 13:20160156. <https://doi.org/10.1098/rsif.2016.0156>
36. Cleveland WS, Grosse E, Shyu WM. Local regression models. In: Chambers JM, Hastie TJ, editors. *Statistical models in S*. London, UK: Chapman & Hall; 1991. p. 309–376.
37. Loader C. *Local Regression and Likelihood*. New York, NY: Springer-Verlag New York; 1999.
38. Hart JD. Automated kernel smoothing of dependent data by using time series cross-validation. *Journal of the Royal Statistical Society B (Statistical Methodology)*. 1994; 56(3):529–542.
39. Gillespie DT. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*. 2007; 58:35–55. <https://doi.org/10.1146/annurev.physchem.58.032806.104637> PMID: 17037977
40. Johnson P. adaptivetau: Tau-leaping stochastic simulation; 2016. Available from: <https://CRAN.R-project.org/package=adaptivetau>.
41. Elaydi S. *An Introduction to Difference Equations*. New York, NY: Springer; 2005.

42. Bauch CT, Earn DJD. Transients and attractors in epidemics. *Proceedings of the Royal Society of London B*. 2003; 270(1524):1573–1578. <https://doi.org/10.1098/rspb.2003.2410>
43. Earn DJD. Mathematical epidemiology of infectious diseases. In: Lewis MA, Chaplain MAJ, Keener JP, Maini PK, editors. *Mathematical biology*. vol. 14 of IAS Park City Mathematics Series. American Mathematical Society; 2009. p. 151–186.
44. Fine PEM. The interval between successive cases of an infectious disease. *American Journal of Epidemiology*. 2003; 158(11):1039–1047. <https://doi.org/10.1093/aje/kwg251> PMID: 14630599