TECHNICAL APPENDIX

Model Equations

The model is formulated as the following system of ordinary differential equations:

$$\begin{aligned}
\frac{dS}{dt} &= \alpha(N-S) - \lambda S \frac{\left(I_P + \rho I_S + I_A\right)}{N} \\
\frac{dE}{dt} &= \lambda S \frac{\left(I_P + \rho I_S + I_A\right)}{N} - \left(\zeta + \alpha\right) E \\
\frac{dI_P}{dt} &= \zeta E - (\kappa + \alpha) I_P \\
\frac{dI_S}{dt} &= p_S \kappa I_P - \left(\gamma_S + \alpha\right) I_S \\
\frac{dI_A}{dt} &= \left(1 - p_S\right) \kappa I_P - \left(\gamma_A + \alpha\right) I_A \\
\frac{dR}{dt} &= \gamma_A I_A + \gamma_S I_S - \alpha R
\end{aligned} \quad , \qquad \text{(Eq A1)}$$

where S denotes susceptible to infection, E exposed to infection (infected) but not yet infectious, I_P infectious but pre-symptomatic, I_S symptomatically infected, I_A asymptomatically infected, and R recovered with lifetime immunity to reinfection. Rates, denoted by Greek letters, and there inferred values are given in Table 2 of the main paper.

The parameter ρ takes a value between 0 and 1 and represents the proportion of infected individuals with symptoms who continue to engage in sexual activity and therefore can continue to transmit infection. In this study we assigned ρ the value 0 on the assumption that the severity of symptoms is such that sexual activity is highly unlikely. This parameter therefore does not appear in Table 2 of the main paper and is not included in the steady state equations below.

In order to reduce the number of parameters that must be fitted to outbreak data we assumed that the recovery rates for asymptomatic and symptomatic infection, (γ_A and γ_S, respectively) are equal. This allowed us to further assume that:

$$I_S / I_A = p_S / \left(1 - p_S\right), \qquad \text{(Eq A2)}$$

where p_S is the proportion of those with pre-symptomatic infection that go on to develop symptomatic infection.

The steady-state solution of this system of ODEs was solved in *Mathematica*® (Wolfram Research, Inc., Champaign, IL, USA) yielding the following:

$$\begin{aligned}
S^* &= \frac{N(\alpha+\gamma)(\alpha+\zeta)(\alpha+\kappa)}{\zeta(\alpha+\gamma+p_S\kappa)\lambda} \\
E^* &= N\alpha\left(\frac{1}{\alpha+\zeta} - \frac{(\alpha+\gamma)(\alpha+\kappa)}{\zeta(\alpha+\gamma+p_S\kappa)\lambda}\right) \\
I_P^* &= N\alpha\left(\frac{\zeta}{(\alpha+\zeta)(\alpha+\kappa)} - \frac{\alpha+\gamma}{(\alpha+\gamma+p_S\kappa)\lambda}\right) \\
I_S^* &= p_S N\alpha\kappa\left(\frac{\zeta}{(\alpha+\gamma)(\alpha+\zeta)(\alpha+\kappa)} - \frac{1}{(\alpha+\gamma+p_S\kappa)\lambda}\right) \\
R^* &= N\gamma\kappa\left(\frac{\zeta}{(\alpha+\gamma)(\alpha+\zeta)(\alpha+\kappa)} - \frac{1}{(\alpha+\gamma+p_S\kappa)\lambda}\right)
\end{aligned} \quad . \qquad \text{(Eq A3)}$$

Under the assumption expressed in Eq A2 above, I_A can be expressed as a function of I_S and therefore does not appear in Eq A3.

Fitting

Poisson Noise

As described in the main text, we assessed robustness by generating 1,000 alternate datasets. These were generated using a parametric bootstrap, where the observed monthly incidence totals were taken as the means of Poisson random variables for each month and new datasets were generated by simultaneously sampling from each month's distribution using the MATLAB routine *poissrnd*. The transmission model was then fitted to each of these revised datasets using the *multistart* routine in MATLAB, with 100 iterations and the optimisation function *lsqcurvefit*. Initial parameter values for each *multistart* iteration were sampled from the predefined fitting constraints given in Table 2 of the main text. In Figure A1, we show the 95% confidence intervals around data points under this procedure, along with the original data and an example bootstrapped epidemic curve. Figure A2 shows the distributions for all model parameters obtained by fitting the model to the simulated datasets, and Figure A3 shows the distribution for the total duration of infection obtained by summing the durations in E, I_P and I_S.

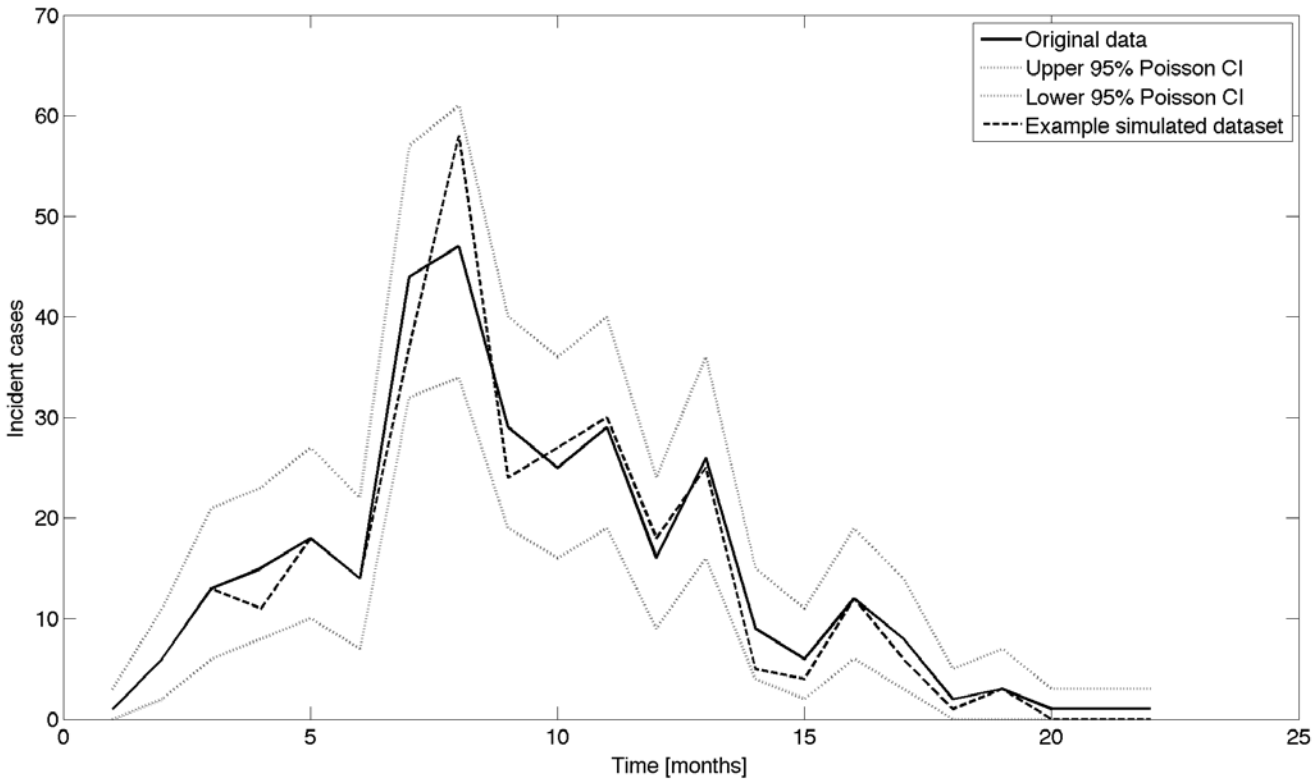

Figure A1. Outbreak data, 95% confidence intervals and example dataset obtained from parametric bootstrap procedure.

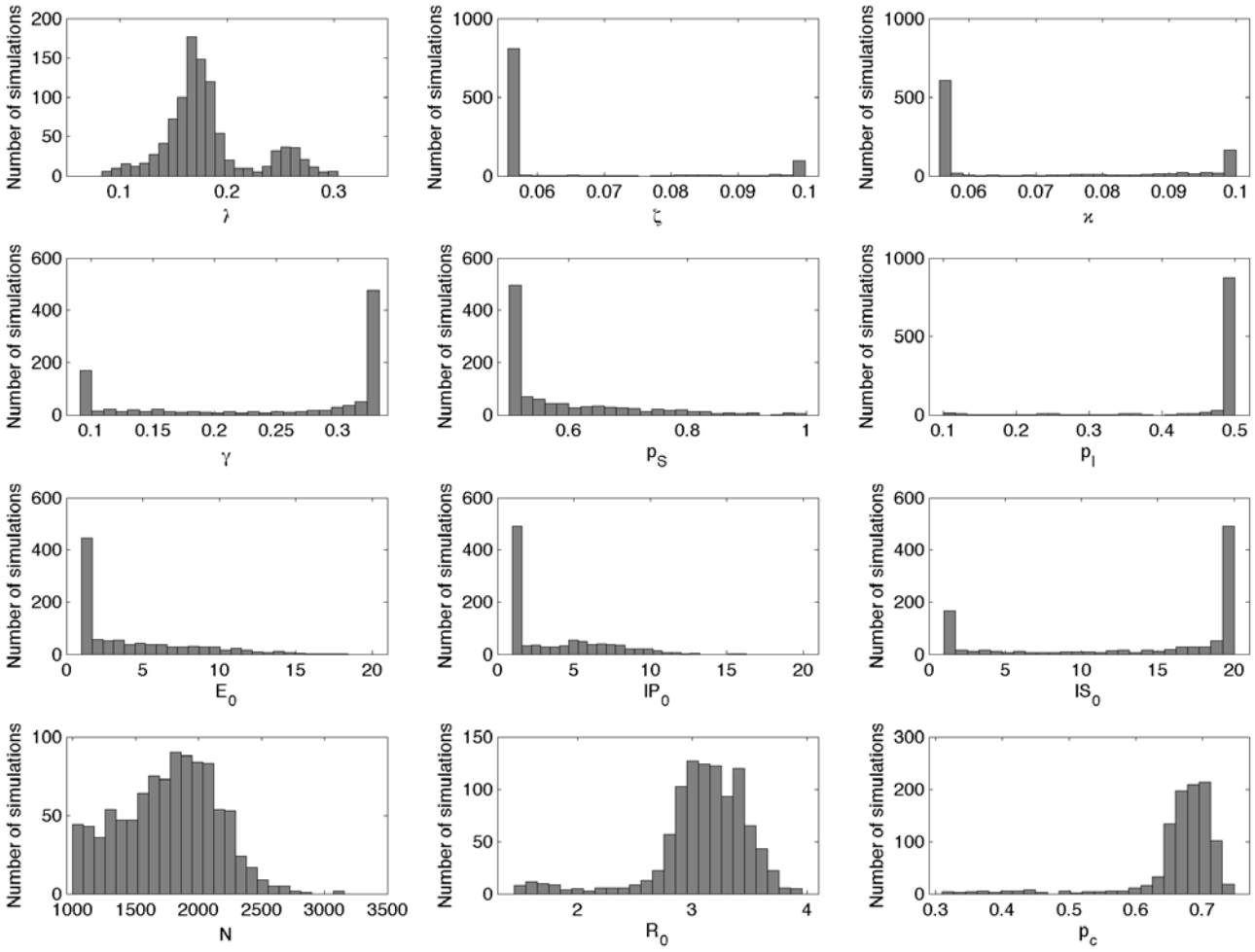

Figure A2. Distributions of model parameters obtained by fitting the model to simulated datasets.

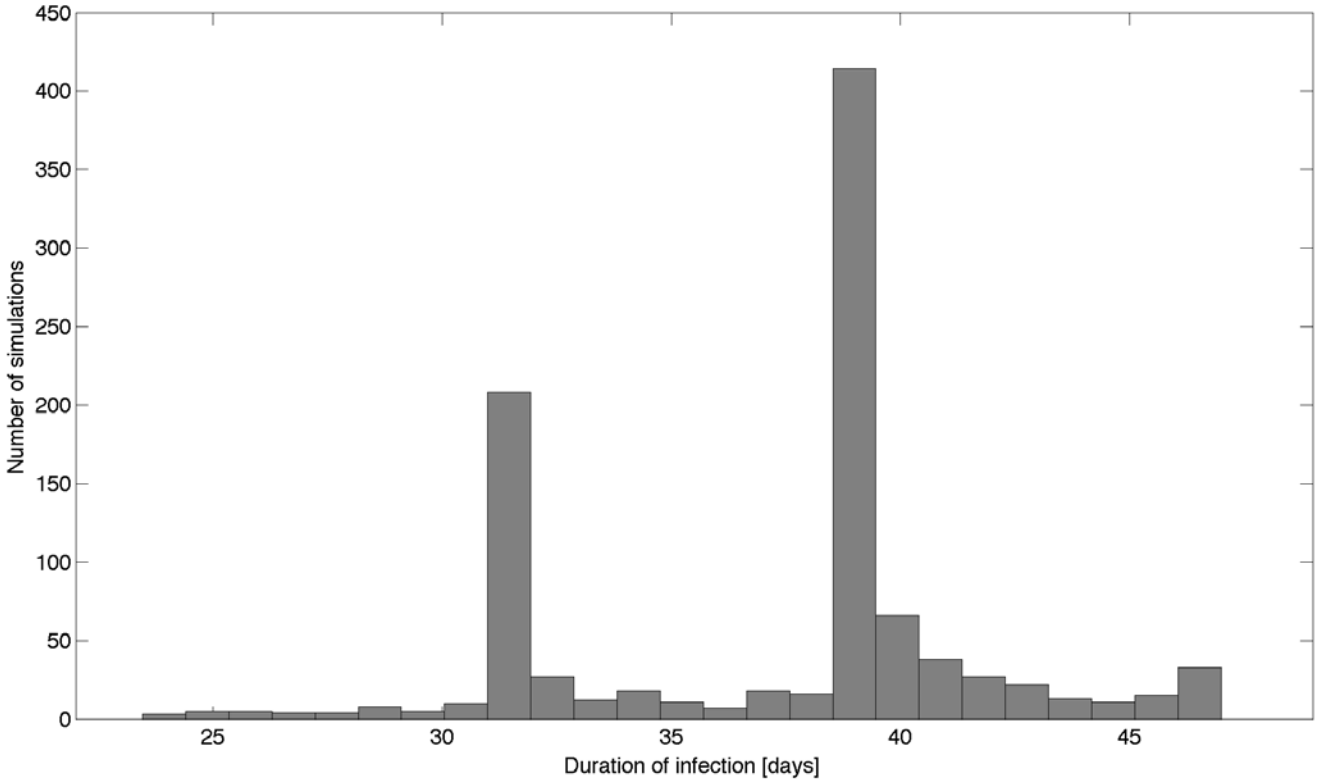

Figure A3. Distribution of total duration of infection obtained by summing the durations in E, I_P and I_S.

Simulated Outbreaks

We used a mixture modelling approach [1] to approximate the simulated distribution of outbreak sizes, with a lognormal component describing the self-limiting outbreaks and a normal component describing outbreaks that "took-off". As the simulated outbreak data was discrete and our model continuous, we first converted our model to a probability mass model, with values given by the cumulative probability between (non-negative) integer outbreak sizes:

$$P(x) = p\int_{x}^{x+1} \frac{1}{t\sigma_l\sqrt{2\pi}} e^{-\frac{(\ln t-\mu_l)^2}{2\sigma_l{}^2}} dt + \frac{(1-p)}{\sigma_n\sqrt{2\pi}} \int_{x}^{x+1} e^{-\frac{(t-\mu_n)^2}{2\sigma_n{}^2}} dt, \quad x \in \mathbb{Z}_+ , \qquad \text{(Eq A4)}$$

where the parameter *p* is the mixture proportion attributed to the log-normal term, μ_l and σ_l are the lognormal mean and standard deviation respectively, while μ_n and σ_n are the normal mean and standard deviation respectively. The log-likelihood for this model is then simply

$$L = \sum_{x=0}^{N} Y(x)\log(P(x)), \qquad \text{(Eq A5)}$$

where $Y(x)$ is the number of simulated outbreaks of size x. Maximum-likelihood fits of this model to the simulated data were computed using the MATLAB routine *fminunc*, which is an unconstrained optimisation procedure. However, we first transformed the variance parameters and mixture parameters by using log and logit transforms respectively to ensure that estimates for variances and the mixture component probability *p* were constrained to positive values and the [0,1] range, respectively. We also added a small noise term (1e-7) to each probability term to avoid log(0) errors. Fits to the simulated data were in general very good but the approach was unable to capture the exact point at which at which the outbreak component should become zero (an immune proportion of 0.747 as derived from the largest R0 estimate). Thus we used the mixture models to determine the outbreak probability only for immune-proportions between 0 and 0.7, with values 0.75 and above taken to be 0 in accordance with the identified herd immunity threshold. Examples of the mixture fits are shown in Figure A4 and illustrate that the mixture probabilities accurately describe the probability of outbreaks occurring. The fits to the lognormal component are uniformly good but while the normal component captures the mean and variance of the data well, it is unable to accommodate the left-skewness of the outbreak component. However, as estimation of the mixture components were the primary focus here, we viewed this as acceptable.

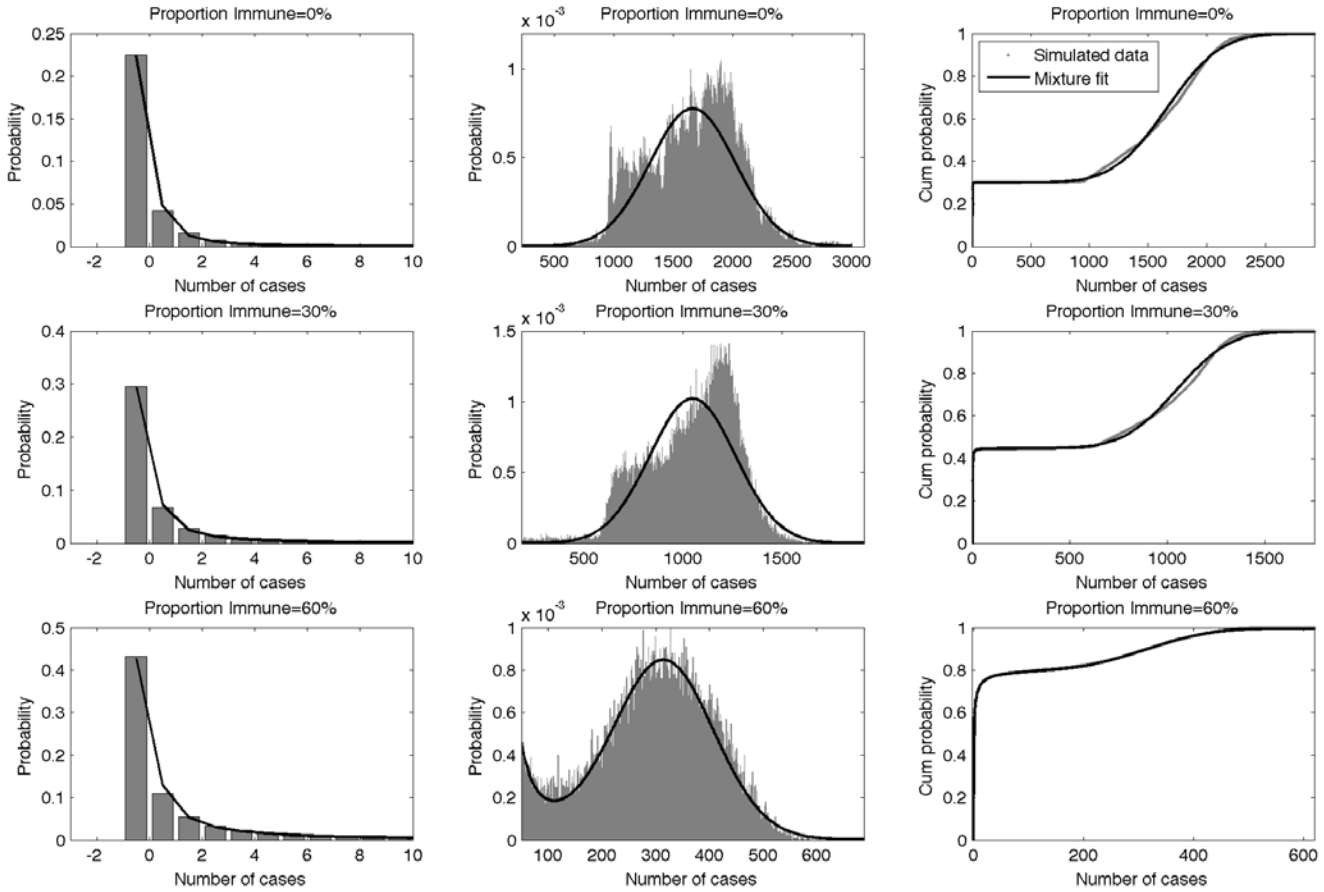

Figure A4: Fits of mixture components, with column 1 showing the lognormal fit to the non-outbreak component and column 2 showing the normal fit to the outbreak component. Column 3 shows the fit of the cumulative distribution to the data, indicating the accuracy of the mixture components. The rows differ by the assumed proportion immune as indicated in panel titles.

Sensitivity Analysis

Partial rank correlation coefficients (PRCC) [2] were calculated to assess the relative importance of the fitted parameters with respect to the goodness of the fit and epidemic potential; mean square error (MSE) and R0 were the respective outcomes in this analysis. SaSAT software [3] was used to generate 10,000 parameter vectors by the method of Latin hypercube sampling (LHS) [4]. Uniform distributions were specified for all parameters with the same upper and lower bounds used in the fitting process. The model was run for each parameter vector and the MSE and R0 calculated for each. PRCCs were then calculated in SaSAT for parameter sets yielding MSE < 500 (n = 2,285), MSE < 300 (n = 681) and MSE < 100 (n = 39). The results are summarised in Table A2 below.

Table A2: PRCCs and importance ranking for model parameters with respect to MSE and R_0

Parameter	MSE < 500		MSE < 300		MSE < 100	
	PRCC (rank)	p-value	PRCC (rank)	p-value	PRCC (rank)	p-value
MSE						
N	-0.085 (9)	<0.001	-0.378 (10)	0.328	0.069 (9)	0.723
λ	-0.838 (1)	<0.001	-0.415 (1)	<0.001	-0.129 (7)	0.504
ζ	0.089 (7)	<0.001	-0.054 (8)	0.160	0.400 (3)	0.032
κ	0.436 (2)	<0.001	0.322 (3)	<0.001	0.559 (1)	0.002
γ	0.114 (5)	<0.001	0.098 (5)	0.011	-0.059 (10)	0.761
p_S	0.155 (4)	<0.001	0.085 (7)	0.028	0.416 (2)	0.025
p_I	0.431 (3)	<0.001	0.363 (2)	<0.001	0.070 (8)	0.719
$E_{initial}$	-0.090 (6)	<0.001	-0.143 (4)	<0.001	0.213 (5)	0.268
$I_{P,initial}$	0.086 (8)	<0.001	-0.047 (9)	0.222	0.395 (4)	0.034
$I_{S,initial}$	-0.001 (10)	0.947	-0.092 (6)	0.017	0.179 (6)	0.354
R_0						
N	0.053 (6)	0.012	-0.213 (6)	<0.001	-0.388 (5)	0.038
λ	0.992 (1)	<0.001	0.874 (1)	<0.001	0.547 (4)	0.002
ζ	-0.018 (7)	0.384	0.040 (8)	0.306	-0.347 (6)	0.065
κ	-0.878 (2)	<0.001	-0.805 (2)	<0.001	-0.678 (1)	<0.001
γ	-0.381 (4)	<0.001	-0.376 (5)	<0.001	-0.322 (7)	0.088
p_S	-0.579 (3)	<0.001	-0.623 (3)	<0.001	-0.573 (3)	0.001
p_I	0.085 (5)	<0.001	0.409 (4)	<0.001	0.638 (2)	<0.001
$E_{initial}$	0.010 (9)	0.633	-0.038 (9)	0.329	-0.305 (8)	0.108
$I_{P,initial}$	0.009 (10)	0.669	-0.040 (7)	0.300	-0.146 (9)	0.449
$I_{S,initial}$	-0.018 (8)	0.385	-0.015 (10)	0.703	-0.056 (10)	0.772

References

(1) McLachlan GJ, Peel D. Finite Mixture Models. New York: Wiley-Interscience, 2000.

(2) Saltelli A, et al. Sensitivity Analysis In Practise: A Guide to Assessing Scientific Models. Chichester: Wiley, 2004.

(3) Hoare A, Regan DG, Wilson DP. Sampling and sensitivity analyses tools (SaSAT) for computational modelling. Theoretical Biology and Medical Modelling 2008; 5: 1-18.

(4) Blower SM, Dowlatabadi H. Sensitivity and uncertainty analysis of complex-models of disease transmission: an HIV model, as an example. International Statistical Review 1994; 62: 229-243.