frontiers
in Applied Mathematics
and Statistics

TYPE Original Research
PUBLISHED 02 April 2026
DOI 10.3389/fams.2026.1798581

OPEN ACCESS

EDITED BY
Lubna Pinky,
Meharry Medical College, United States

REVIEWED BY
Gilberto Gonzalez-Parra,
New Mexico Tech, United States
You Chang,
University of Copenhagen, Denmark

*CORRESPONDENCE
Renata Retkute
✉ rr614@cam.ac.uk

RECEIVED 28 January 2026
REVISED 21 February 2026
ACCEPTED 09 March 2026
PUBLISHED 02 April 2026

CITATION
Retkute R, Hollingsworth TD and Minter A (2026) Formulating likelihood functions for infectious disease dynamics for neglected tropical diseases. *Front. Appl. Math. Stat.* 12:1798581. doi: 10.3389/fams.2026.1798581

Formulating likelihood functions for infectious disease dynamics for neglected tropical diseases

Renata Retkute[1]*, T. Déirdre Hollingsworth[2] and Amanda Minter[2]

[1]Epidemiology and Modelling Group, Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom, [2]Big Data Institute, Li Ka Shing Centre for Health, Information and Discovery, University of Oxford, Oxford, United Kingdom

Reliable inference in infectious disease modeling requires careful treatment of both model structure and the relationship between latent infection dynamics and observed data. Likelihood functions, which link model parameters to empirical observations, can be formulated either to explicitly represent underlying disease transmission and reporting processes (process-based) or to summarize statistical patterns in aggregated outcomes (observation-based). Stochastic models capture inherent variability in transmission and detection, whereas deterministic models describe average system behavior and often rely on statistical assumptions to account for residual uncertainty. Using two neglected tropical disease (NTD) models, we compare parameter estimation based on complete individual-level events with that based on aggregated counts. By generating synthetic outbreak data from stochastic simulations and analyzing it under alternative modeling frameworks, we show how different combinations of model formulation and likelihood structure influence both point estimates and uncertainty quantification. Our findings indicate that, even when detailed process information is unavailable, observation-based likelihoods can produce robust parameter estimates and credible uncertainty intervals, highlighting their usefulness for practical decision-making in contexts with limited or aggregated surveillance data.

KEYWORDS
leprosy, likelihood, neglected tropical diseases, trachoma eradication, transmission model

1 Introduction

A central challenge in epidemiological research is calibrating infectious disease models to observed data, integrating different modeling approaches to quantify and propagate uncertainty through inference and prediction. Likelihood functions, which formalize the probability of observing data given a set of model parameters, serve as the critical interface linking model structure, stochastic processes, and observed data, and thus underpin both frequentist and Bayesian inference. Model parameters can be estimated using maximum likelihood estimation (MLE) [1], or within Bayesian frameworks using methods such as Markov Chain Monte Carlo (MCMC) [2, 3] or importance sampling [4, 5], while likelihood-free approaches such as Approximate Bayesian Computation (ABC) provide practical alternatives when the likelihood is analytically intractable

[6–8]. Across these approaches, the specific formulation of the likelihood—ranging from process-based models that explicitly simulate both disease transmission and reporting processes, to observation-based models that rely on statistical assumptions about measured outcomes—determines how uncertainty arising from stochasticity, data limitations, and model structure is captured and propagated. This integration of modeling methods is particularly important for neglected tropical diseases (NTDs), where dynamic transmission models are increasingly used to evaluate progress toward WHO 2030 Roadmap targets [9, 10], and where reliable parameter inference is essential not only for predicting intervention impacts but also for estimating the true burden of infection, providing decision-makers with quantitatively rigorous uncertainty estimates to guide policy and control strategies.

In a Bayesian framework, the posterior distribution represents the updated uncertainty about model parameters after observing data. It is defined as the normalized product of the prior distribution and the likelihood [11]. The formulation of the likelihood is therefore central to how uncertainty from both the disease process and the observation process is propagated into parameter estimates and predictive outputs. For disease transmission models, likelihoods can be constructed in several ways. *Process-based likelihoods* (PBL) explicitly simulate both the latent disease states and, where applicable, the reporting process, capturing stochasticity inherent in transmission, recovery, and case detection; a notable example is partially observed Markov process models applied to human Ebola outbreaks, which integrate unobserved infection dynamics with observed case counts. *Observation-based likelihoods* (OBL), in contrast, assume that the data follow a statistical distribution around model outputs, such as Normal [1, 12], Poisson [13], Student's t [14], negative binomial [15, 16], or beta-binomial [17]. Gaussian likelihoods correspond to a least-squares error framework, while other distributions accommodate count or overdispersed data. Breto [18] distinguished mechanistic likelihoods reflecting either the transmission or the measurement process, underscoring how process- and observation-based formulations represent sources of uncertainty differently.

Closely linked to likelihood formulation is the choice between stochastic and deterministic model structures. Stochastic models naturally incorporate variability in infection and reporting processes, while deterministic models approximate mean system behavior, requiring observation-based likelihoods to capture residual uncertainty [19]. Thus, integrating model structure, likelihood choice, and stochastic representation is essential for quantifying uncertainty in a coherent and interpretable manner, particularly in policy-relevant applications where predictive confidence intervals (CIs) guide intervention decisions [20].

To illustrate these concepts, we formulate likelihood functions for compartmental models using both process-based and observation-based approaches. We generate synthetic data with a stochastic model and aggregate individual events into population-level longitudinal observations to mimic real-world surveillance. Parameter inference is then performed under two scenarios: one using a stochastic model with a process-based likelihood, and another using a deterministic model with an observation-based likelihood applied to aggregated data. We apply this framework to two case studies: trachoma and leprosy. Trachoma, caused by *Chlamydia trachomatis*, remains a leading cause of infectious blindness worldwide and exhibits complex, seasonal transmission patterns [21]. Leprosy is a chronic bacterial disease with slow progression and low incidence, making its transmission dynamics difficult to observe directly [22]. These contrasting diseases provide complementary contexts to evaluate how different likelihood formulations and model structures perform under low-incidence vs. more aggregated, population-level transmission scenarios. We systematically compare how different likelihood formulations and model structures affect parameter estimation and uncertainty quantification. These case studies show how observation-based likelihoods can yield reliable parameter estimates and uncertainty quantification in situations where data are aggregated, as is typical in many real-world surveillance settings.

2 Materials and methods

2.1 Process-based likelihood

We assume that events can arise in the time period $[t, t+dt]$, where dt is a vanishingly small time period, with the probability that an event occurs given in Equation 1:

$$P\left(\text{event in interval } [t, t+dt]\right) = \lambda(t)dt, \tag{1}$$

where $\lambda(t)$ is a rate function. If we further assume that the probabilities of events in distinct intervals are independent, it follows that for any interval $(t, t+u)$ [23] the probability of no events is given by Equation 2:

$$P\left(\text{no events in interval } (t, t+u)\right) = \exp\left(-\int_t^{t+u} \lambda(\tau)d\tau\right). \tag{2}$$

We order events into a series of non-decreasing times: $t_{(0)} = 0 < t_{(1)} < \ldots < t_{(j)} < t_J$, where t_J is the time of last observed event.

The process-based likelihood can be obtained by multiplying: (i) likelihood of an event occurring at time $t_{(j)}$; (ii) likelihood that no events occur in the interval $(t_{(j-1)}, t_{(j)})$. Then the likelihood is as in Equation 3 with:

$$\mathcal{L}^P(\text{events}) = \prod_{j=1}^{J}\left[\lambda(t_{(j)})\exp\left(-\int_{t_{(j-1)}}^{t_{(j)}} \lambda(\tau)d\tau\right)\right]. \tag{3}$$

For compartmental disease transmission models, the rate $\lambda(t)$ is piece-wise constant and can change only at event times. This means that the integral in Equation 3 is equal to the value of the rate function at the time of previous event multiplied by the length of time to the next event, i.e. $\int_x^y \lambda(\tau)d\tau = \lambda(x) \times (y-x)$.

2.2 Observation based likelihood

We assume the data are available at the population level across multiple observation points. Observation-based likelihood can be formulated by assuming that observed data are normal-distributed [1], Poisson-distributed [13, 24], Student's t-distributed

[14], negative binomial-distributed [15, 16], or beta-binomial-distributed [17] around an output of a model.

2.3 Case studies

2.3.1 Trachoma transmission model

A simple susceptible, infected, and susceptible (SIS) model of ocular chlamydial infection has been proposed to model transmission in a core group of children [25, 26]. In this framework, all individuals in a population are classified according to their epidemiological status over time. We assume a constant population of size N, ignoring births and deaths, and that individuals mix homogeneously. These assumptions are reasonable for modeling short-term infection dynamics in a closed population such as a school or village.

In the SIS model, individuals move between susceptible (S) and infected (I) states, with $S_t = N - I_t$ at any time t. We can define changes in the number of infected individuals by Equation 4:

$$\frac{dI_t}{dt} = \beta\,(N - I_t)\,\frac{I_t}{N} - \gamma I_t, \tag{4}$$

where I_t is the number of infectious cases at time t, β is the transmission rate, and γ is the rate of recovery from infection. The initial number of infected individuals, I_0, is specified based on the observed data or a small seed value, and $S_0 = N - I_0$.

The observation-based likelihood can be obtained by assuming that the observed number of infected cases (I^{obs}) is Poisson-distributed around a solution of Equation 4 (I^{sim}):

$$\mathcal{L}^{O}(I^{obs} * d * d = 0, t_{obs}|\beta,\gamma) = \prod_{d=0}^{t_{obs}} P_{Poiss}\left(I_d^{obs}|I_d^{sim}(\beta,\gamma)\right). \tag{5}$$

We chose a Poisson likelihood because it is appropriate for count data and does not require estimating additional dispersion parameters, unlike Gaussian or negative binomial likelihoods. It assumes independence of observations given the model. In practice, overdispersion may occur, but Poisson provides a simple baseline.

Based on Equation 4, we can define a stochastic model. In the stochastic formulation, infection and recovery events occur probabilistically in a small time interval dt. This can be implemented using the Gillespie stochastic simulation algorithm or an equivalent discrete-time approximation. The probabilities that an event occurs in the time period $[t, t + dt]$ are:

$$\begin{aligned} P(S \to I) &= \beta S_t I_t N^{-1} dt, \\ P(I \to S) &= \gamma I_t dt, \end{aligned} \tag{6}$$

At each step, the model selects whether an infection or recovery occurs based on these probabilities. The stochastic model reflects the same dynamics as the deterministic ODE but accounts for random fluctuations in small populations.

The process-based likelihood can be obtained by multiplying: (i) the likelihood that either infection or recovery occurred at each observed event time $t_{(j)}$ and no events occurred between consecutive observation times $(t_{(j-1)}, t_{(j)})$; and (ii) the likelihood of no events occurring between the last event and the end of the observational window:

$$\mathcal{L}^{P}(\{t_{(j)}, \eta_{(j)}\}_{j=0,J}|\beta,\gamma) = \left\{\prod_{j=1}^{J}\left[\underbrace{\left(\frac{\beta}{N}S_{t_{(j-1)}}I_{t_{(j-1)}}\right)^{\eta_{(j)}}}_{\text{Infection}}\underbrace{\left(\gamma I_{t_{(j-1)}}\right)^{1-\eta_{(j)}}}_{\text{Recovery}}\right]\underbrace{\exp\left[-\left(\frac{\beta}{N}S_{t_{(j-1)}}I_{t_{(j-1)}}\right)\left(t_{(j)} - t_{(j-1)}\right)\right]}_{\text{No infection}}\underbrace{\exp\left[-\left(\gamma I_{t_{(j-1)}}\right)\left(t_{(j)} - t_{(j-1)}\right)\right]}_{\text{No recovery}}\right\}$$

Here we ordered observed events into a series of times of either infection or recovery: $t_{(0)} = 0 < t_{(1)} < \ldots < t_{(k)} < t_J$. For each event, we introduced an index of event, η, where $\eta_{(j)} = 1$ if event at time $t_{(j)}$ is an infection, and zero if it is recovery [6]. This formulation captures the probability of observing the exact sequence of events under the stochastic model and is widely used in survival and counting-process frameworks.

2.3.2 Leprosy-like transmission model

We simulated a disease trajectory using a general model that describes the effect of passive surveillance on the prevention of epidemics of neglected tropical diseases [27]. The model is described by a system of ordinary differential equations in which individuals are classified as either susceptible to infection (S), infected but not yet infectious (E), infected and infectious (I), detected (D), or recovered (R).

Individuals in the I and D classes contribute to onward infection at a rate β. Newly infected individuals become infectious at a rate σ. A proportion of these individuals, p_E, move directly to the detected class (D), and the remaining proportion remains undetected and moves to the infectious class (I).

Individuals in the detected class can either receive treatment at a rate ρ or recover at a rate γ. Infectious individuals can also recover at the same rate γ. Demographic processes are included, individuals are assumed to be born susceptible, and all individuals can die at a natural mortality rate μ.

$$\frac{dS}{dt} = \mu N - \beta S(I + D)N^{-1} - \mu S, \tag{7}$$

$$\frac{dE}{dt} = \beta S(I + D)N^{-1} - \sigma E - \mu E, \tag{8}$$

$$\frac{dI}{dt} = (1 - p_E)\sigma E - \gamma I - \mu I, \tag{9}$$

$$\frac{dR}{dt} = \gamma I + \gamma D + \rho D - \mu R \tag{10}$$

$$\frac{dD}{dt} = (1 - p_E)\sigma E - \gamma D - \rho D - \mu D \tag{11}$$

10.3389/fams.2026.1798581

The probabilities that an event can occur in time period $[t, t+dt]$ are given by Equations 12–20 with:

$$P(S \to E) = \beta S(I+D)N^{-1}dt, \quad (12)$$
$$P(E \to D) = p_E \sigma E dt, \quad (13)$$
$$P(E \to I) = (1-p_E)\sigma E dt, \quad (14)$$
$$P(I \to R) = \gamma I dt, \quad (15)$$
$$P(D \to R) = \rho D dt, \quad (16)$$
$$P(E \to \varnothing) = \mu E \quad (17)$$
$$P(I \to \varnothing) = \mu I \quad (18)$$
$$P(D \to \varnothing) = \mu D \quad (19)$$
$$P(R \to \varnothing) = \mu R \quad (20)$$

where $\varnothing$ is death. As we assume a closed population, each death leads to the birth of a susceptible.

The process-based likelihood function would be given by:

$$\mathcal{L}^P(\{t_j, \eta_{1(j)}, \eta_{2(j)}, \eta_{3(j)}, \eta_{4(j)}\}_{j=0,obs}|\beta, p_E, \mu, \gamma, \sigma, \rho) =$$
$$\prod_{j=1}^{k}\Bigg[\underbrace{\left(\frac{\beta}{N}S_{t_{(j-1)}}\left(I_{t_{(j-1)}}+D_{t_{(j-1)}}\right)\right)^{\eta_{1(j)}(1-\eta_{4(j)})}}_{\text{Exposure}}$$
$$\underbrace{Bern\left(\eta_{2(j)}, p_E\right)^{\eta_{1(j)}}}_{\text{Detected or infected}}$$
$$\underbrace{\left(\sigma E_{t_{(j-1)}}\right)^{(1-\eta_{1(j)})(1-\eta_{3(j)})(1-\eta_{4(j)})}}_{\text{Infection}}$$
$$\underbrace{\left(\gamma I_{t_{(j-1)}}\right)^{(1-\eta_{1(j)})(1-\eta_{2(j)})\eta_{3(j)}(1-\eta_{4(i)})}}_{\text{Recovery of infected}}$$
$$\underbrace{\left((\gamma+\rho)\, D_{t_{(j-1)}}\right)^{(1-\eta_{1(j)})(1-\eta_{2(j)})(1-\eta_{3(j)})(1-\eta_{4(i)})}}_{\text{Recovery of detected}}$$
$$\underbrace{\left(\mu E_{t_{(j-1)}}\right)^{\eta_{1(j)}(1-\eta_{2(j)})(1-\eta_{3(i)})\eta_{4(i)}}}_{\text{Death of exposed}}$$
$$\underbrace{\left(\mu I_{t_{(j-1)}}\right)^{(1-\eta_{1(j)})(1-\eta_{2(j)})(1-\eta_{3(i)})\eta_{4(i)}}}_{\text{Death of infected}}$$
$$\underbrace{\left(\mu D_{t_{(j-1)}}\right)^{(1-\eta_{1(j)})\eta_{2(j)}(1-\eta_{3(i)})\eta_{4(i)}}}_{\text{Death of detected}}$$
$$\underbrace{\left(\mu R_{t_{(j-1)}}\right)^{(1-\eta_{1(j)})(1-\eta_{2(j)})\eta_{3(i)}\eta_{4(i)}}}_{\text{Death of removed}}\Bigg]$$
$$\underbrace{\exp\left[-\left(\frac{\beta}{N}S_{t_{(j-1)}}\left(I_{t_{(j-1)}}+D_{t_{(j-1)}}\right)\right)\left(t_{(j)}-t_{(j-1)}\right)\right]}_{\text{No infection}}$$
$$\underbrace{\exp\left[-\left(\sigma E_{t_{(j-1)}}\right)\left(t_{(j)}-t_{(j-1)}\right)\right]}_{\text{No exposure}}$$
$$\underbrace{\exp\left[-\left(\gamma I_{t_{(j-1)}}+(\gamma+\rho)D_{t_{(j-1)}}\right)\left(t_{(j)}-t_{(j-1)}\right)\right]}_{\text{No recovery}}$$
$$\underbrace{\exp\left[-\left(\mu\left(E_{t_{(j-1)}}+I_{t_{(j-1)}}+D_{t_{(j-1)}}+R_{t_{(j-1)}}\right)\right)\left(t_{(j)}-t_{(j-1)}\right)\right]}_{\text{No death}}, \quad (21)$$

where $Bern(x, p_E)$ is the density function for the Bernoulli distribution with probability of success on each trial equal to p_E. We introduced a set of binary indicator variables: $\eta_{1(i)} = 1$ if event $t_{(i)}$ is an infection, and 0 otherwise; $\eta_{2(i)} = 1$ if exposed becomes detected, 0 if infectious; $\eta_{3(i)} = 1$ if an event is recovery, 0 otherwise; $\eta_{4(i)} = 1$ if an event is death, 0 otherwise.

2.4 Comparing estimation methods

We compared infection-process-based and observation-process-based likelihood formulations using the Akaike Information Criterion (AIC). For a model with a log likelihood $\mathcal{L}(\theta)$ and k independently estimated parameters, AIC is defined as

$$\text{AIC}(\theta) = -2\log\mathcal{L}(\theta) + 2k, \quad (22)$$

where θ denotes parameter values and k the number of fitted parameters [28]. AIC provides an information-theoretic measure of relative model support by balancing goodness-of-fit against model complexity [29]. Parameter vectors with lower AIC are considered to have a better balance of fit [30].

To examine parameter identifiability and the geometry of the supported parameter space, we evaluated AIC over a grid of parameter values. Two-dimensional AIC surfaces were constructed to visualize relationships between parameters.

Inference was restricted to parameter combinations satisfying

$$\mathcal{B}_i = \{i : \text{AIC}(\theta_i) - \min(\text{AIC}) \le 2\}, \quad (23)$$

which defines the set of parameters with substantial empirical support [31]. Parameter combinations exceeding this threshold were considered unsupported and were masked in surface plots to enhance interpretability.

3 Results

3.1 Trachoma transmission

We have set population size to $N = 100$, transmission rate $\beta = 0.044$ and recovery rate $\gamma = 0.017$ [25]. We simulated outbreaks using Gillespie's stochastic simulation algorithm [32]. The times of individual events are shown in Figure 1a, with an arbitrary vertical scale showing a cumulative increase in the number of events. For process-based parameter estimation, we assumed that we know the type and timing of all events shown in Figure 1a. This is an idealized situation, and would be possible only under

10.3389/fams.2026.1798581

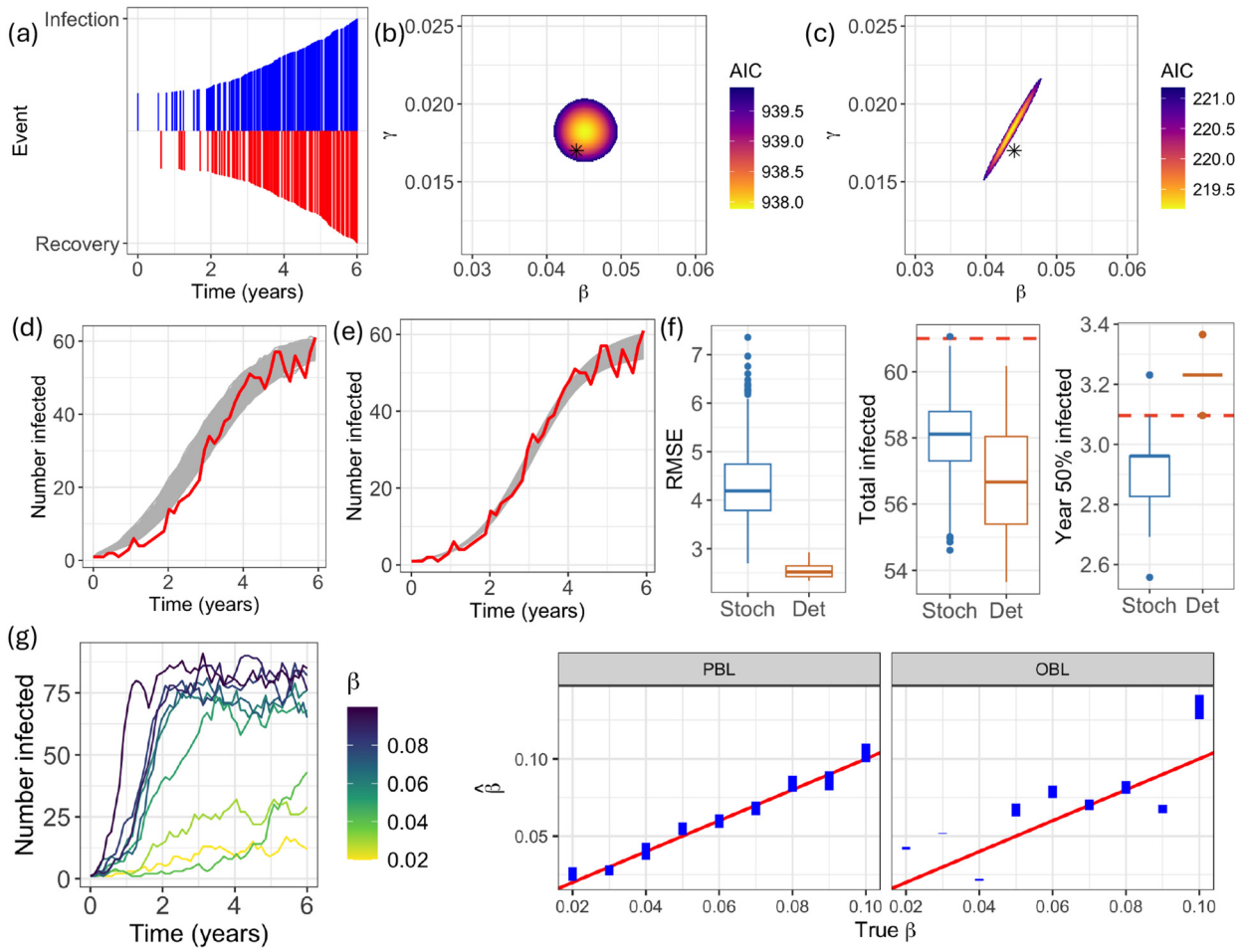

FIGURE 1
Comparing estimation methods for trachoma transmission. **(a)** Times of individual events (infection and recovery). **(b)** AIC surface for the PBL. **(c)** AIC surface for the OBL, calculated using the weekly number of infected individuals. In **(b, c)**, colors indicate parameter combinations within [min(AIC), min(AIC) + 2]. Black stars denote the true parameter values ($\beta = 0.044$, $\gamma = 0.017$). **(d)** Posterior simulations of the number of infections (gray lines) compared with observed values (red line), based on the stochastic model and the PBL. Each curve represents the average of 100 simulations for a given parameter set. **(e)** Posterior simulations of the number of infections (gray lines) compared with observed values (red line), based on the deterministic model and the OBL. **(f)** Distributions of summary statistics: Root Mean Square Error (RMSE), the number of infected individuals at $t = 6$ years, and the year in which the outbreak reached 50% of the maximum number of infections. Horizontal red dashed lines indicate the values of these summary statistics for the observed data. **(g)** Sensitivity analysis examining the impact of different input parameter values on parameter estimation. Simulated observation data for $\beta \in [0.02, 0.1]$ (left panel) and the distribution of estimated ranges for $\hat{\gamma}$ (right panel).

controlled conditions in a laboratory, with constant monitoring of symptoms or frequent testing. For OBL parameter estimation, we used the weekly number of infectious individuals obtained from the stochastic model as our observed data.

Figures 1b, c display the AIC surfaces obtained under the process-based likelihood (PBL) and observation-based likelihood (OBL) formulations. Both approaches exhibit a similar qualitative structure but differ in their ability to recover the true parameter values. In neither case are the true parameter combinations located at the center of the region of strongest empirical support, indicating some degree of estimation bias.

Under the PBL formulation, the true parameter values lie within the conventional support region defined by $\min(\text{AIC}) \leq \text{AIC} \leq \min(\text{AIC}) + 2$, and are therefore statistically competitive with the best-fitting parameter set. In contrast, under the OBL formulation, the true parameter values fall outside the $\Delta\text{AIC} \leq 2$ region, indicating weaker empirical support relative to the AIC-optimal estimate. Differences between the two methods are also reflected in the estimated recovery rate. For OBL, the supported range is $\hat{\gamma} \in [0.016, 0.020]$, whereas for PBL it is $\hat{\gamma} \in [0.015, 0.021]$. Although these intervals are close in magnitude, the dependence structure between parameters differs markedly. Under OBL, the estimated transmission and recovery rates exhibit a strong positive correlation ($\rho = 0.985$). By contrast, under PBL the correlation is negligible ($\rho = -0.001$), indicating improved parameter separability.

To assess whether these discrepancies in parameter recovery translate into differences in model performance, we conducted simulation experiments using all parameter sets that satisfied the acceptance criterion defined in Equation 23. For each admissible parameter set, we generated model outputs and compared the simulated weekly incidence with the observed weekly number

of infected individuals. For the stochastic transmission model, 100 independent realizations were performed per parameter set to capture intrinsic variability, and the mean weekly number of infections across these simulations was used for comparison with the empirical data. As expected, both methods produced model trajectories that closely matched the observed data (Figures 1d, e). Here, the OBL outcome showed less variability and lay closer to the observed data at the early stage of the outbreak (the first four years) than the PBL trajectories.

To quantify goodness of fit, we calculated three summary statistics: the Root Mean Square Error (RMSE), the number of infected individuals at $t = 6$ years, and the year in which the outbreak reached 50% of the maximum number of infections (Figure 1f). RMSE differed substantially between estimation methods. The deterministic approach (Det) produced a markedly lower mean RMSE (mean = 2.54, 95% CI: 2.53–2.55) compared with the stochastic approach (Stoch) (mean = 4.30, 95% CI: 4.24–4.36). Variability was also considerably smaller for the deterministic method (SD = 0.140) than for the stochastic method (SD = 0.745), indicating greater stability in predictive performance. The number of infected individuals at $t = 6$ years differed slightly between estimation methods. The stochastic approach (Stoch) produced a higher mean number of infections (mean = 58.0, 95% CI: 57.9–58.1) compared with the deterministic approach (Det) (mean = 56.7, 95% CI: 56.6–56.8). Variability was slightly larger for the deterministic method (SD = 1.65) than for the stochastic method (SD = 1.10). Despite this difference, both methods underpredicted the observed number of infections at 6 years, which was 61, indicating that neither model fully captured the outbreak magnitude at this time point. The year in which the outbreak reached 50% of the maximum number of infections differed between estimation methods. The stochastic approach (Stoch) predicted this milestone slightly earlier (mean = 2.94 years, 95% CI: 2.93–2.95) compared with the deterministic approach (Det) (mean = 3.21 years, 95% CI: 3.20–3.21). Overall, these results indicate that while PBL recovers parameters more accurately, OBL provides better predictive trajectories in the presence of noisy data, highlighting a trade-off between parameter recovery and model performance.

We evaluated the robustness of parameter recovery by varying the true transmission rate ($\beta \in [0.02, 0.10]$) in simulated datasets and estimating parameters using the PBL and OBL (Figure 1g). For each value of β, confidence bounds were defined by the AIC support region $[\min(\text{AIC}), \min(\text{AIC}) + 2]$. Under the PBL, the true transmission rate consistently fell within the estimated support intervals across the full range of simulated β values. The estimated ranges were generally centered on the true values, with moderate and relatively stable interval widths, indicating good practical identifiability and limited bias across low- and high-transmission settings.

In contrast, the OBL displayed substantial variability and systematic bias in several scenarios. At higher transmission intensities ($\beta = 0.10$), the OBL overestimated β, with support intervals shifted above the true value. At intermediate levels, estimates were occasionally downward biased (e.g., $\beta = 0.09$). In contrast, at lower transmission intensities ($\beta \leq 0.04$), the method produced either markedly underestimated or overestimated ranges, often with narrow intervals that did not include the true parameter.

Models can be used to assess the effect of treatment programs on disease elimination, for example, the mass treatment of trachoma with antibiotics. Here, we investigated how parameter estimates derived from process-based vs. observation-based likelihoods influence model predictions of the time required until elimination. We simulated treatment by reducing the number of infected individuals, assuming 80% effective coverage and bi-annual administration, as in Ray et al. [25], with MDA starting at year 6. Elimination was defined as $I < 1$, and models were run using parameter sets satisfying the condition given by Equation 23. Control of trachoma via bi-annual MDA over a 3.5-year period is illustrated in Figure 2. Figure 2a shows an example of prevalence dynamics simulated using the stochastic model, while Figure 2b shows the corresponding dynamics under the deterministic model. In both Figures 2a, b, green vertical lines indicate the timing of MDA applications. These simulations highlight the impact of repeated treatment on reducing prevalence over time and allow comparison of variability between stochastic and deterministic model predictions. We found that both models predicted successful trachoma elimination (Figure 2c). The expected time to elimination was 2.5–3 years in the deterministic model and 2.2–4.1 years in the stochastic model. In the stochastic formulation, the variability in time to elimination across posterior simulations reflects the effect of parameter uncertainty, and the expected time until elimination was positively correlated with pre-treatment trachoma prevalence ($\rho = 0.73$). These observations are in agreement with [25], where both deterministic and stochastic model simulations showed that prevalence of infection is progressively reduced with each periodic treatment.

3.2 Leprosy-like transmission

We have set population size to $N = 1000$, transmission rate $\beta = 0.1$, recovery rate $\gamma = 0.017$, and proportion of detected individuals to $p_E = 0.01$ [27]. Simulated epidemiological curves for both deterministic and stochastic formulations are shown in Figures 3a, b. Again, for PBL parameter estimation, we assumed that we know the type and timing of all events (including different types of infections, recoveries, deaths, and births). However, for OBL parameter estimation, we used only a weekly number of detected cases (inset in Figure 3b). Non-zero values in the detected compartment (D) occurred between approximately years 2 and 4. The number of detected individuals remained low throughout this period, with a maximum of 3 cases observed. We chose the Poisson distribution for observation-based likelihood.

Parameter estimation using the PBL approach recovered ranges that included the true transmission rate and were moderately wide (Figure 3c). Specifically, β was estimated between 0.091 and 0.101, encompassing the true value of 0.100, while the detection probability p_E was estimated between 0.006 and 0.016, also covering the true value of 0.010, albeit with greater relative uncertainty. Overall, PBL produced reliable intervals that captured the true parameter values, particularly for p_E, despite being broader. In contrast, the OBL approach yielded substantially narrower ranges (Figure 3d). The transmission rate was estimated between

10.3389/fams.2026.1798581

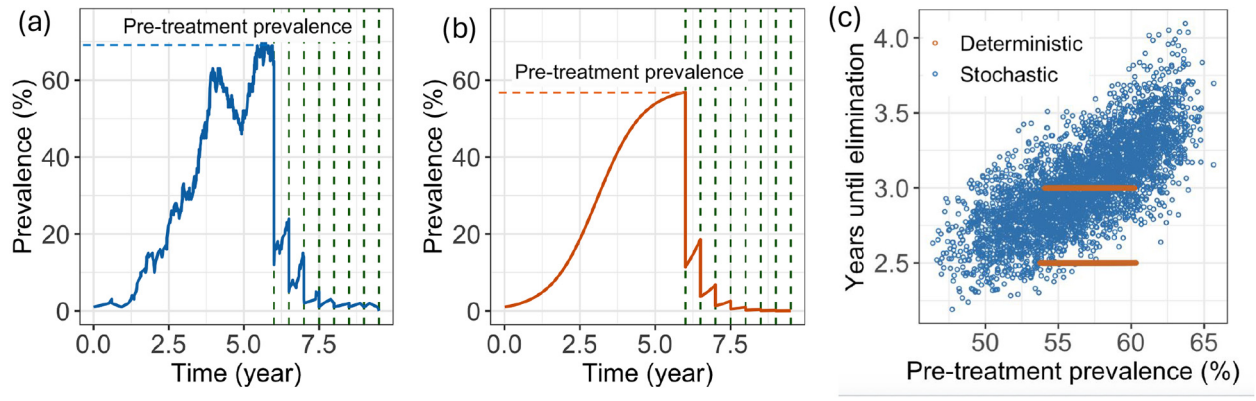

FIGURE 2
Control of trachoma via bi-annual MDA after 6 years. **(a)** Example of prevalence dynamics based on a stochastic model. **(b)** Example of prevalence dynamics based on a deterministic model. In **(a, b)**, green vertical lines show the application of MDA. **(c)** Predicted years of bi-annual control until elimination of trachoma.

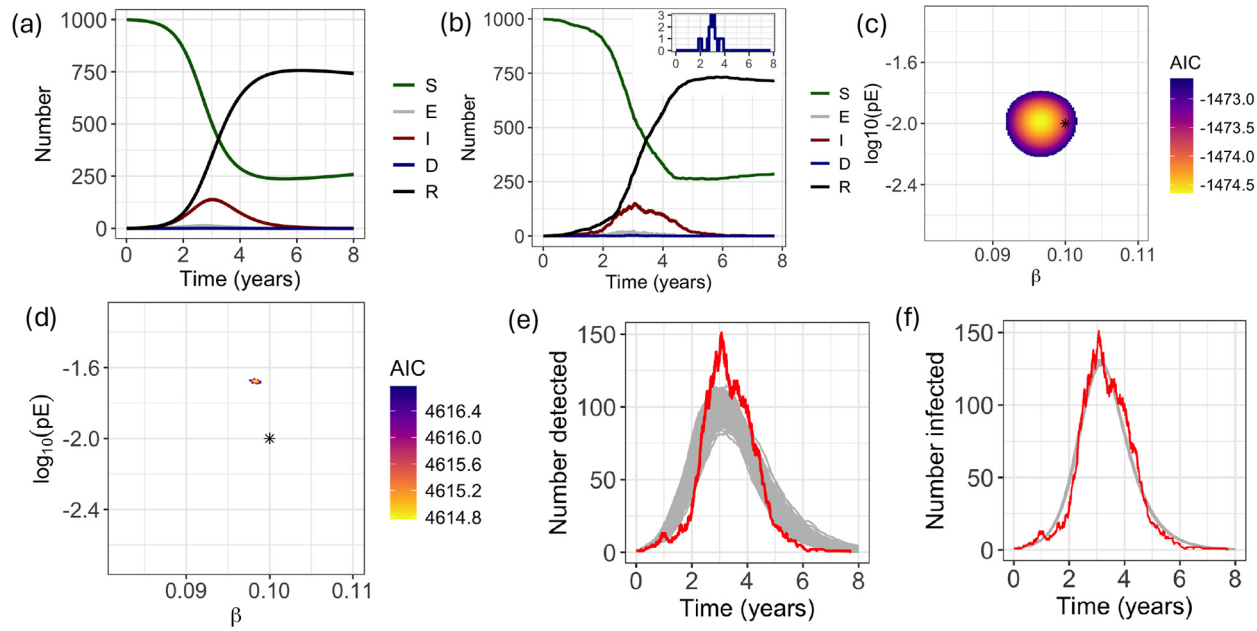

FIGURE 3
Simulated outbreak of leprosy and log-likelihood surfaces. **(a)** epidemiological curves for deterministic model; **(b)** epidemiological curves for stochastic model (insert shows a close up on several detected cases); **(c)** AIC surface for the PBL; **(d)** AIC surface for the OBL calculated using weekly number of detected. In **(c, d)** colors indicate parameter combinations within [min(AIC), min(AIC) + 2]. Black stars show the true values of parameters ($\beta = 0.1$, p_E=0.01). **(e)** Posterior simulations of the number of infected (gray lines) compared with observed values (red line), based on the stochastic model and the PBL. Each curve represents the average of 100 simulations for a given parameter set. **(f)** Posterior simulations of the number of infected (gray lines) compared with observed values (red line), based on the deterministic model and the OBL.

0.097 and 0.098, narrowly missing the true value, while p_E was estimated between 0.020 and 0.021, approximately twice the true value and not overlapping with it. Correlation between β and p_E was negligible for PBL ($\rho = 0.002$) but moderately negative for OBL ($\rho = -0.316$), reflecting a modest trade-off between parameters under observation-based inference. Thus, OBL provided higher precision at the cost of bias, particularly for the detection probability, whereas PBL produced wider but more trustworthy intervals. Consistent with the trachoma case study, the deterministic model combined with OBL showed better agreement with observed data compared with the stochastic model and PBL (Figures 3e, f), even when using only limited information on detected cases.

4 Discussion

In this study, we brought together two complementary approaches to likelihood formulation for infectious disease models: process-based likelihoods and observation-based likelihoods. Although the choice of likelihood is usually dictated by the

type of data available, our objective was to assess how these alternative formulations differ in their treatment of uncertainty when applied to the same underlying outbreak but observed at different levels of granularity. While many NTD models are individual-based models [33, 34], we illustrated our examples using deterministic ordinary differential equation models and their equivalent stochastic versions, allowing us to disentangle the effects of model structure from those of likelihood specification. We based our analysis on two case studies: a simulated outbreak of trachoma in children, and a simulated outbreak based on a leprosy-like transmission model. For both case studies, we used data on all events when calculating process-based likelihoods (time and event type). In contrast, for the observation-based likelihood formulation, we restricted our knowledge to aggregated data from a single compartment: for trachoma, weekly counts of infectious individuals; and for the leprosy-like disease, weekly counts of detected individuals. This design enabled a controlled comparison of how information loss through aggregation affects parameter inference and uncertainty representation.

We performed two-dimensional visualizations of the AIC surface and compared the parameter sets within the support region (defined as $AIC \leq \min(AIC) + 2$) with the true values. For models with more than two parameters, the same approach can be extended by examining pairwise AIC surfaces conditioned on fixed values of additional parameters, or by visualizing three-dimensional AIC contours, to assess identifiability and parameter trade-offs in higher dimensions. Exploring the AIC surface provides insight into identifiability, curvature, and potential sources of bias, beyond what is apparent from point estimates alone. Both frameworks produced reasonable parameter estimates. However, biases can still arise even when all processes are observed, and the likelihood is correctly specified, highlighting the intrinsic uncertainty associated with stochastic disease dynamics. For example, in the leprosy simulations, the OBL produced narrower but biased estimates of the detection probability p_E. In contrast, the PBL produced wider intervals that captured the true value (Figures 3c, d). Similarly, for trachoma, the AIC surfaces revealed that OBL estimates were systematically shifted relative to PBL, reflecting information loss due to temporal aggregation (Figures 1b, c).

The differences in parameter uncertainty between PBL and OBL formulations can be understood mechanistically through the lens of identifiability and information content. In the PBL framework, the full sequence of event times and types constrains the joint parameter space because each parameter contributes to distinct events. For example, infection and recovery events directly inform the relative magnitudes of transmission and recovery rates, improving structural identifiability and reducing parameter correlation. By contrast, under OBL, the data are temporally aggregated counts from a single compartment. This aggregation induces information loss: multiple underlying transmission and detection trajectories can generate similar observed summaries. In particular, the strong positive correlation between transmission and recovery rates under OBL reflects a structural trade-off: increases in transmission can be offset by increases in recovery, yielding similar prevalence trajectories. The PBL formulation mitigates this effect because the timing of individual infection and recovery events breaks this symmetry, allowing separate identification of transition rates. Thus, the narrower yet biased intervals observed under OBL in some scenarios do not necessarily indicate stronger identifiability, but rather reflect the projection of a higher-dimensional stochastic process onto a lower-dimensional summary statistic. However, posterior simulations based on PBL parameters showed much greater variability in outbreak trajectories and worse summary statistics overall, highlighting a practical trade-off between mechanistic identifiability and predictive stability.

Mathematical models are important tools for guiding control and elimination strategies for NTDs. Investigation of the leprosy-like transmission model showed that a high proportion of new infections would need to be detected to prevent resurgence [27]. A crucial question is determining what proportion of infected individuals self-report after symptom onset. Our results indicate that this epidemiologically critical parameter can be inferred not only from full individual-level process information, but also from aggregated surveillance data using observation-based likelihoods, with comparable levels of uncertainty. For the trachoma case study, we found that the two models and likelihood formulations agreed on the feasibility of mass drug administration (MDA) for disease elimination, suggesting that policy-relevant conclusions may be robust to different modeling and likelihood choices when uncertainty is appropriately accounted for.

More broadly, our findings suggest that observation-based likelihoods represent a more realistic and practical framework for infectious disease inference in the majority of the applied settings. In routine surveillance systems, individual-level event histories are rarely available; instead, data typically consist of aggregated case counts collected at regular intervals. The OBL formulation aligns naturally with this data structure and therefore reflects the level of information that is genuinely obtainable in public health contexts. In addition, OBL is substantially simpler to implement computationally, as it avoids explicit reconstruction of latent event histories and reduces the dimensionality of the likelihood evaluation. This can lead to improved numerical stability and lower computational cost, particularly when scaling to large populations or extended time horizons.

Although OBL may introduce modest biases or altered uncertainty structures due to information loss and parameter trade-offs, these disadvantages are generally outweighed by its realism and feasibility. In practical applications, the availability, quality, and granularity of data impose fundamental constraints on inference. A likelihood framework that matches these constraints is therefore preferable, even if it sacrifices some theoretical precision. Taken together, our results indicate that observation-based likelihoods provide a robust, implementable, and policy-relevant approach for parameter estimation in infectious disease models, with minor limitations relative to their substantial practical advantages.

5 Conclusion

Accurate parameter estimation is essential for setting policy targets, guiding surveillance, identifying hotspots [4, 10], designing surveys [27, 35], and communicating uncertainty to decision-makers [36, 37]. A central challenge in achieving these goals is deciding how to represent and propagate uncertainty when data are limited, aggregated, or noisy. By directly comparing

stochastic and deterministic model formulations alongside process-based and observation-based likelihoods, we show that reliable inference does not always require full individual-level data or fully stochastic models. In practice, full event-level data are most critical in situations where parameters are tightly coupled, rare events dominate the dynamics, or the system exhibits strong stochasticity—such as low-incidence outbreaks, diseases with highly heterogeneous transmission, or when multiple unobserved compartments contribute substantially to observed outcomes [38]. Nevertheless, carefully constructed observation-based likelihoods applied to aggregated data can recover key parameters with comparable accuracy and provide meaningful characterization of uncertainty. These findings support an integrated modeling perspective, in which model structure and likelihood formulation are jointly considered as tools for uncertainty quantification rather than viewed as purely technical choices. They provide practical guidance for infectious disease modelers working on NTDs and other data-constrained settings, and align with the broader aim of the special issue to advance transparent, uncertainty-aware infectious disease modeling.

Data availability statement

The datasets and code for this study can be found in the Gtihub repository https://github.com/rretkute/likelihood-functions-ntd.

Author contributions

RR: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. TH: Funding acquisition, Writing – original draft, Writing – review & editing. AM: Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. RR acknowledges support through an award from the BBSRC Flexible Talent Mobility Account. AM and TH acknowledge the Bill and Melinda Gates Foundation via the NTD Modelling Consortium (grant number INV-030046).

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

1. Ionides E, Breto C, King A. Inference for nonlinear dynamical systems. *PNAS*. (2006) 103:18438–43. doi: 10.1073/pnas.0603181103

2. Robert CP. *The Bayesian Choice: Decision-Theoretic Foundations to Computational Implementation*. New York: Springer. (2007).

3. Jewell CP, Kypraios T, Neal P, Roberts GO. Bayesian analysis for emerging infectious diseases. *Bayesian Anal*. (2009) 4:465–96. doi: 10.1214/09-BA417

4. Retkute R, Touloupou P, Basáñez MG, Hollingsworth TD, Spencer SEF. Integrating geostatistical maps and infectious disease transmission models using adaptive multiple importance sampling. *Ann Appl Statist*. (2021) 15:1486. doi: 10.1214/21-AOAS1486

5. Retkute R, Gilligan CA. A novel two-stage parameter estimation framework integrating approximate bayesian computation and machine learning: the ABC-RF-rejection algorithm. *arXiv preprint arXiv:2507.02072* (2025).

6. McKinley T, Cook AR, Deardon R. Inference in epidemic models without likelihoods. *Int J Biostatist*. (2009) 5:24. doi: 10.2202/1557-4679.1171

7. Minter A, Retkute R. Approximate Bayesian Computation for infectious disease modelling. *Epidemics*. (2019) 29:100368. doi: 10.1016/j.epidem.2019.100368

8. Li X, Chadwick F, Swallow B. Advances in approximate Bayesian inference for models in epidemiology. *Epidemics*. (2025) 53:100855. doi: 10.1016/j.epidem.2025.100855

9. World Health Organization (WHO). *Ending the neglect to attain the sustainable development goals: a road map for neglected tropical diseases 2021–2030* (2020). Available online at: https://apps.who.int/iris/bitstream/handle/10665/332094/WHO-UCN-NTD-2020.01-eng.pdf (Accessed June 19, 2023).

10. Vasconcelos A, Nunes-Alves C, Hollingsworth TD. New tools and nuanced interventions to accelerate achievement of the 2030 roadmap for neglected tropical diseases. *Clin Infect Dis*. (2024) 78:S77–82. doi: 10.1093/cid/ciae070

11. O'Neill P. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathem Biosci*. (2002) 180:103–114. doi: 10.1016/S0025-5564(02)00109-8

12. Yang W, Karspeck A, Shaman J. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS Comput Biol*. (2014) 10:e1003583. doi: 10.1371/journal.pcbi.1003583

13. Bootsma MCJ, Ferguson NM. The effect of public health measures on the 1918 influenza pandemic in US cities. *PNAS*. (2007) 104:7588–93. doi: 10.1073/pnas.0611071104

14. Dehning J, Zierenberg J, Spitzner FP, Wibral M, Pinheiro Neto J, Wilczek M, et al. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science*. (2020) 369:160. doi: 10.1126/science.abb9789

15. Cauchemez S, Ferguson NM. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *J R Soc Interface*. (2008) 5:885–97. doi: 10.1098/rsif.2007.1292

16. Stocks T, Britton T, Höhle M. Model selection and parameter estimation for dynamic epidemic models via iterated filtering: application to rotavirus in Germany. *Biostatistics*. (2020) 21:400–16. doi: 10.1093/biostatistics/kxy057

17. Knock ES, Whittles LK, Lees JA, Perez-Guzman PN, Verity R, FitzJohn RG, et al. Key epidemiological drivers and impact of interventions in the 2020 SARS-CoV-2 epidemic in England. *Sci Transl Med*. (2021) 13:eabg4262. doi: 10.1126/scitranslmed.abg4262

18. Breto C. Modeling and inference for infectious disease dynamics: a likelihood-based approach. *Statist Sci*. (2018) 33:57. doi: 10.1214/17-STS636

19. Champagne C, Cazelles B. Comparison of stochastic and deterministic frameworks in dengue modelling. *Math Biosci*. (2019) 310:1–12. doi: 10.1016/j.mbs.2019.01.010

20. Flaig J, Houy N. Disease X epidemic control using a stochastic model and a deterministic approximation: performance comparison with and without parameter uncertainties. *Comput Methods Programs Biomed*. (2024) 249:108136. doi: 10.1016/j.cmpb.2024.108136

21. Mariotti SP, Pascolini D, Rose-Nussbaumer J. Trachoma: global magnitude of a preventable cause of blindness. *Br J Ophthalmol*. (2008) 93:563–8. doi: 10.1136/bjo.2008.148494

22. Richardus JH, Habbema JDF. The impact of leprosy control on the transmission of *M. leprae*: is elimination being attained? *Leprosy Rev*. (2007) 78:330–7. doi: 10.47276/lr.78.4.330

23. Ross S. *Simulation*. Orlando: Academic Press. (2006).

24. Read JM, Bridgen JRE, Cummings DAT, Ho A, Jewell CP. Novel coronavirus 2019-nCoV (COVID-19): early estimation of epidemiological parameters and epidemic size estimates. *Philos Trans R Soc B*. (2021) 376:20200265. doi: 10.1098/rstb.2020.0265

25. Ray KJ, Porco TC, Hong KC, Lee DC, Alemayehu W, Melese M, et al. A rationale for continuing mass antibiotic distributions for trachoma. *BMC Infect Dis*. (2007) 7:91. doi: 10.1186/1471-2334-7-91

26. Blumberg S, Prada JM, Tedijanto C, Deiner MS, Godwin WW, Emerson PM, et al. Forecasting trachoma control and identifying transmission-hotspots. *Clin Infect Dis*. (2021) 72:S134–9. doi: 10.1093/cid/ciab189

27. Minter A, Medley GF, Hollingsworth TD. Using passive surveillance to maintain elimination as a public health problem for neglected tropical diseases: a model-based exploration. *Clin Infect Dis*. (2024) 78:S169–74. doi: 10.1093/cid/ciae097

28. Akaike H. Information theory and an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory* (1973). p. 267–281.

29. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer. (2002).

30. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr*. (1974) 19:716–23. doi: 10.1109/TAC.1974.1100705

31. Chang Y, de Jong MCM. A novel method to jointly estimate transmission rate and decay rate parameters in environmental transmission models. *Epidemics*. (2023) 42:100672. doi: 10.1016/j.epidem.2023.100672

32. Gillespie DT. Approximate accelerated stochastic simulation of chemically reacting systems. *J Chem Phys*. (2001) 115:1716–1733. doi: 10.1063/1.1378322

33. Anderson RM, Basáñez MG, Sturrock RJ. Mathematical models for neglected tropical diseases: Essential tools for control and elimination, part A: Volume 87. In: *Advances in parasitology*. San Diego, CA: Academic Press. (2015).

34. Anderson RM, Basáñez MG, Sturrock RJ. Mathematical models for neglected tropical diseases: Essential tools for control and elimination, part B: Volume 94. In: *Advances in parasitology*. San Diego, CA: Academic Press. (2016).

35. Diggle PJ, Fronterre C, Gass K, Hundley L, Niles-Robin R, Sampson A, et al. Modernizing the design and analysis of prevalence surveys for neglected tropical diseases. *Philos Trans R Soc B*. (2023) 378:20220276. doi: 10.1098/rstb.2022.0276

36. Behrend MR. Basáñez MG, Hamley JID, Porco TC, Stolk WA, Walker M, et al. Modelling for policy: the five principles of the neglected tropical diseases modelling consortium. *PLoS Negl Trop Dis*. (2020) 14:e0008033. doi: 10.1371/journal.pntd.0008033

37. Bergström F, Favero M, Britton T. Identifiability in epidemic models with prior immunity and under-reporting. *arXiv preprint arXiv:2506.07825* (2025).

38. Dankwa EA, Brouwer AF, Donnelly CA. Structural identifiability of compartmental models for infectious disease transmission is influenced by data type. *Epidemics*. (2022) 41:100643. doi: 10.1016/j.epidem.2022.100643