Estimating seroconversion rates accounting for repeated infections by approximate Bayesian computation

Peter Teunis[1], Yuke Wang[1], Kristen Aiemjoy[2,3],
Mirjam Kretzschmar[4,5], Marc Aerts[6]

October 11, 2023

Corresponding author:
PFM Teunis PhD
Center for Global Safe WASH, Hubert Department of Global Health,
Emory University Rollins School of Public Health
1518 Clifton Rd. NE, Atlanta, GA 30322 USA
email: peter.teunis@emory.edu

1. Center for Global Safe WASH, Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA

2. Division of Epidemiology, Department of Public Health Sciences, University of California, Davis, CA,USA

3. Department of Microbiology and Immunology, Mahidol University Faculty of Tropical Medicine, Bangkok, Thailand

4. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

5. Center for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

6. Center for Statistics (CenStat), University Hasselt, Belgium

Abstract

This study presents a novel approach for inferring the incidence of infections by employing a quantitative model of the serum antibody response. Current methodologies often overlook the cumulative effect of an individual's infection history, making it challenging to obtain a marginal distribution for antibody concentrations.

Our proposed approach leverages approximate Bayesian computation (ABC) to simulate cross-sectional antibody responses and compare these to observed data, factoring in the impact of repeated infections. We then assess the empirical distribution functions of the simulated and observed antibody data utilizing Kolmogorov deviance, thereby incorporating a goodness–of–fit check. This new method not only matches the computational efficiency of preceding likelihood-based analyses but also facilitates the joint estimation of antibody noise parameters.

The results affirm that the predictions generated by our within–host model closely align with the observed distributions from cross–sectional samples of a well–characterized population. Our findings mirror those of likelihood–based methodologies in scenarios of low infection pressure, such as the transmission of pertussis in Europe. However, our simulations reveal that in settings of higher infection pressure, likelihood–based approaches tend to underestimate the force of infection. Thus, our novel methodology presents significant advancements in estimating infection incidence, thereby enhancing our understanding of disease dynamics in the field of epidemiology.

Introduction

Serum antibodies specific to a particular pathogen not only act as a historical marker for past infection, but elevated concentrations may also suggest a recent encounter with the same pathogen and possibly, associated protective immunity[1]. In clinical settings high antibody levels have been commonly characterized as exceeding a threshold or cutoff level to identify recently infected subjects. In sero–epidemiology the same approach is used for classifying seropositive and seronegative subjects, but often with a somewhat lower cutoff, to capture less recent infections[2,3,4]. Using a fixed cutoff ignores variation in seroresponses between individuals and variation in time since the most recent infection[1]. As current serum antibody levels carry information about the infection history of the sampled persons, it is possible to not just determine the prevalence of infection but also infer the seroincidence: the rate with which seroconversions occur in the study population[5].

Most studies using serology to estimate incidence rely on directly detecting seroconversions[6,7], or using age profiles of seroprevalence to infer incidence[8,9]. More recently, machine learning methods have been employed to exploit the kinetics of the seroresponse[10].

The serum antibody response to infection can be described quantitatively by a within–host model, relating a transient increase in antigen present to antibody mediated pathogen removal or inactivation. When the time course of the seroresponse to infection is known quantitatively, any antibody concentration measured in a population sample may be translated into a time when the most recent infection occurred. This is not completely straightforward because infections occur randomly, and people are also sampled randomly, independent of their infection history. Moreover, there is strong variation in seroresponses among individuals. A high antibody concentration in a cross–sectional sample implies that infection occurred recently, but also that the sample came from a person who seroconverted to a high concentration. A low antibody concentration may indicate that the sampled subject seroconverted a while ago, but it is also possible that that person seroconverted to a low antibody concentration not so long ago. Thus, low and high antibody concentrations contribute different information to measurement of the seroincidence. Such a backcalculation approach[11] is feasible but requires prior assumptions about the distribution of seroincidence due to the uncertainty associated with low cross–sectional antibody levels[12]. A simpler and computationally efficient approach to estimate seroincidence assumed a Poisson infection rate for incident infections to find the marginal distribution of antibody concentrations in a cross–sectional population sample[5]. Allowance could be made for variable infection rates: estimates of the seroincidence could be obtained for small cross–sectional samples sizes ($N < 50$) and variation of incidence among subsets of population data could be studied[13]. Over time refinements were added, accounting for non–exponential antibody decay, age at first infection and antibody noise[14].

The within–host model proposed by de Graaf et al.[15,16] appeared to easily fit a variety of longitudinal serological data[17,13,18,19]. This model allows for adaptative responses because the current response to infection depends on the baseline antibody level, just before the infection event. There is also a threshold for the baseline antibody concentration, above which the subsequent seroresponse makes a “small” jump, surmised to correspond to a mild, asymptomatic infection[15] when the previous infection occurred recently.

Current estimates of seroincidence have ignored the influence of elevated baseline levels on repeated infections. When the infection pressure is high enough so that a person may have experienced multiple infections at the age of sampling, the current seroresponse not only depends on the most recent infection, but on the complete infection history of that person. Although it could be shown that a marginal antibody distribution exists for an exposed population[20], a closed expression that can be plugged into a likelihood function is not available. High infection pressures may however occur during outbreaks or in endemic situations where a single individual may have many infections over a lifetime[21]. High baseline serum antibody concentrations at the time of infection may be associated with (partial) immunity or asymptomatic infections. Therefore it is desirable to be able to apply the within–host model for repeated infections in seroincidence calculations[14]. The present paper takes a new and more flexible approach, simulating cross–sectional antibody distributions and comparing these with observed population samples.

Seroresponse to infection

The within–host model assumes that exposure is followed by an infection phase where antibody concentrations increase exponentially and pathogen multiplication is inhibited proportionally to circulating antibody concentrations[15]. At the time pathogens are cleared, the (net) antibody production is downregulated, beginning a phase of prolonged antibody decay[16]. More details are given in the Appendix (A.1). The time course of the seroresponse then becomes:

$$y(\tau) = \begin{cases} y_0 \mathrm{e}^{\mu_1 \tau} & 0 \leq \tau \leq t_1 \\ y_1 \left(1 + (r-1) y_1^{r-1} \alpha (\tau - t_1)\right)^{-\frac{1}{r-1}} & t_1 \leq \tau \end{cases} . \tag{1}$$

Where y_0 is the baseline antibody concentration. At $\tau = 0$ the subject is exposed and at $\tau = t_1$ pathogens are cleared. As antibody levels increase during infection and decrease after infection, at t_1 antibody concentrations are at a maximum y_1. Antibody decay is described by a power function with parameters α (rate) and r (shape) allowing for non–exponential shapes, indicative of within–host heterogeneity in antibody production[16].

Suppose the rate parameters μ_0 (net growth rate of pathogens), μ_1 (net rate of increase of serum antibodies), and c ("efficacy" of antibodies in removing pathogens) remain fixed for successive infections in the same host. Then the baseline antibody level at time of infection determines the next peak antibody level[15]

$$y_1 = y_0 \left(1 + \frac{(\mu_1 - \mu_0) b_0}{c y_0}\right)^{\frac{\mu_1}{\mu_1 - \mu_0}}, \tag{2}$$

where b_0 is the initial pathogen concentration. When the antibody concentration at the time of infection exceeds a threshold

$$y_0 > y_{\text{min}} = \frac{\mu_0 b_0}{c}, \tag{3}$$

the pathogen concentration decreases monotonically and the antibody concentration makes a "small jump" (Fig. 3 in de Graaf et al.[15]) corresponding to a mild (presumably asymptomatic) infection. When longitudinal antibody data are available for a cohort of infected subjects the model parameters $(\mu_0, \mu_1, c^* = c/b_0)$ may be estimated, defining the relation between baseline y_0 and the following peak y_1 (eq. 2), estimates shown in Figure A3 in the Appendix.

Figure 1 here.

At the time of the first infection, some time after birth, antibody concentrations are presumed to be low (ignoring short–lived maternal immunity) and the first seroresponse tends to reach a high peak level, corresponding to a “large jump”, likely symptomatic[15]. When the next infection happens to occur not too much later, antibody concentrations likely have not returned to low levels and the new infection may be milder, corresponding to a “small jump”, presumably asymptomatic. Figure 1 shows the time course of pertussis (IgG–PT) antibody levels during the life of a hypothetical person, with a sequence of symptomatic and asymptomatic infections.

Estimating seroconversion rates

In a population exposed to an infectious pathogen, individuals seroconvert with a certain frequency, dependent on the infection pressure. Serum antibodies measured in a random sample from such an exposed population have a distribution that depends upon the infection rate λ: when λ is low there is a high chance of sampling from individuals who were never infected or were infected a long time ago, leading to low antibody levels. When λ is high there will be more individuals with high antibody levels, as the probability of a recent infection increases.

With known kinetics of the seroresponse, and assuming that infections occur as a Poisson process, antibody levels in a cross–sectional sample can be used to estimate the infection rate λ for the sampled population[5]. It is possible that a sampled subject has not been infected at all: when a sample has been collected at a young age or when the infection rate is low the probability cannot be neglected that a sampled subject has escaped infection. This profoundly influences the estimation of incidences from cross–sectional population sample data[14]. Moreover, antibody measurements are prone to noise: measurement noise associated with the high sensitivity of assays, but also noise caused by antibodies in the blood sample that were not elicited by the ongoing seroresponse but that do react with the used antigens (termed M–

noise and B–noise respectively[14]).

In assessing the likelihood for cross-sectional data the main problem is how to deal with past events: an individual may not have been infected at all; they may have had a single infection or they may have experienced two or more infections during their lifetime. Previous analyses[14] dealt with only two categories (0, and 1 or more infections). Only the most recent infection was accounted for, and the influence of antibodies remaining from previous infections was ignored. Inclusion of one or more infections prior to the most recent one complicates the calculation of the resulting antibody distribution[20] to the extent that likelihood analysis becomes cumbersome.

Fitting based on simulation

In contrast, simulation of serum antibody levels following repeated infections is straighforward, see e.g. Figure 1. Thus, one can generate a simulated cross-sectional sample by repeatedly simulating a life history of seroresponses, ending at a given age at which the subject is sampled. Simulated ages can be matched with the ages of subjects in a population sample to obtain comparable simulation results. Now one would like to know how well antibody levels in the simulated sample match with the observed antibody levels in the population sample.

For a sample of N_{obs} observed antibody concentrations and another sample of N_{sim} simulated antibody concentrations the empirical distribution functions (EDFs)

$$F_{\text{obs}}(y) = \frac{1}{N_{\text{obs}}} \sum_{i=1}^{N_{\text{obs}}} [Y_{\text{obs},i} \leq y] \quad \text{and} \quad F_{\text{sim}}(y) = \frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} [Y_{\text{sim},i} \leq y] \tag{4}$$

can be used to assess the similarity of the observed and simulated samples. Three different classes of statistics for measuring the distance between EDFs have been explored:

1. Based on the positive and negative differences

$$\begin{aligned} D_+ &= \max_y \left(F_{\text{obs}}(y) - F_{\text{sim}}(y)\right) \\ D_- &= \max_y \left(F_{\text{sim}}(y) - F_{\text{obs}}(y)\right) \end{aligned} \tag{5}$$

the (two sample) Kolmogorov–Smirnov (KS) test statistic[22]

$$D_{\text{KS}} = \max(D_+, D_-), \tag{6}$$

may be calculated. D_{KS} can be directly translated into a probability p_{KS}[23].

Alternatively, the Kuiper (KP) test statistic may be calculated[24]

$$V = D_+ + D_-, \tag{7}$$

providing higher sensitivity to differences in the tails of the EDFs.

2. On the other hand, Cramér–von Mises and related tests measure the distance between distribution functions as

$$N_{\text{obs}} \int \left(F_{\text{obs}}(y) - F_{\text{sim}}(y)\right)^2 w(y) \mathrm{d}F_{\text{sim}}(y), \tag{8}$$

where the weighting function $w(y) = 1/\left(F_{\text{sim}}(y)(1 - F_{\text{sim}}(y))\right)$ defines the Anderson–Darling (AD) test statistic which can be calculated as[25]

$$A^2 = -N_{\text{obs}} - \sum_{i=1}^{N_{\text{obs}}} \frac{2i-1}{N_{\text{obs}}} \left(\log(F_{\text{sim}}(Y_{\text{obs},i})) + \log(1 - F_{\text{sim}}(Y_{\text{obs},N_{\text{obs}}+1-i}))\right), \tag{9}$$

for ordered data $Y_{\text{obs},1} < Y_{\text{obs},2} < \cdots < Y_{\text{obs},N_{\text{obs}}}$.

It should be noted that ties (in observed or simulated data) are not considered because these will be unlikely in these (real number) data.

3. The Kullback–Leibler (KL) divergence for the two samples $\boldsymbol{Y}_{\text{obs}}$ and

$\boldsymbol{Y}_{\text{sim}}$ can be approximated by

$$D_{\text{KL}}(Y_{\text{obs}}||Y_{\text{sim}}) = \sum_{i=2}^{N_{\text{obs}}} \log\left(\frac{N_{\text{obs}}(Y_{\text{sim},j} - Y_{\text{sim},k})}{N_{\text{sim}}(Y_{\text{obs},i} - Y_{\text{obs},i-1})}\right), \tag{10}$$

where j and k are chosen such that

$$\max(Y_{\text{sim},j}) \leq Y_{\text{obs},i} < \min(Y_{\text{sim},k}), \tag{11}$$

using the estimator proposed by Perez–Cruz[26]. As above, please note that ties (in observed or simulated data) are not accounted for.

Since calculation of these statistics requires mostly sorting and addition or subtraction, they are simple to implement and can be calculated at high speed. They are therefore well suited to use for estimation requiring repeated evaluations.

Implementation

Simulation of a cross–sectional sample of antibody concentrations for any subject in a population with λ infections per unit time starts with choosing an age a at which a blood sample is taken. For the sampled subject a set of kinetic parameters $(\mu_0, \mu_1, c^*, \alpha, r)$ and initial state y_0 are assigned by random selection from the posterior predictive samples from the longitudinal model, as described in Teunis et al.[16].

Step 1 Then an interval is sampled from an exponential distribution with rate parameter λ: if the interval is longer than the age of the subject, a random sample from a (lognormal) B–noise distribution (parameters μ, σ) is returned as Y_{sim}. When the interval is shorter than the age of sampling, the end of the interval is the age at which the first infection occurs.

Step 2 Now the next interval is sampled, and checked against the remaining time until age of sampling. The seroresponse model is then employed to calculate the antibody concentration at the end of the current infection period (either the next infection or the age of sampling), using eqns. A.2 and A.3 (in the Appendix) and then eqn. 1. When the endpoint is the age of sampling, a random sample from the B–noise distribution is added to the antibody concentration and the result is returned as the current Y_{sim}.

Step 3 When the age of sampling has not been reached, the antibody concentration at the endpoint is used as the new baseline y_0 and the procedure is repeated from Step 2.

Noise is only added at the end, when the antibody level at sampling time has been calculated. It is assumed that B–noise represents antibodies that react with the assay antigen, but are not involved in the ongoing seroresponse, possibly because of cross–reactivity with related pathogens.

The above procedure is repeated for each individual needed for the population sample. When simulating for the purpose of fitting to observed data, the simulated ages can be matched with the ages in the observed population sample.

In case the baseline y_0 at time of infection is to be ignored, only the last interval is needed (the most recent interval that ends with taking a blood sample). The antibody concentration at age of sampling is calculated from the duration of the last interval (step above), with the parameter vector $(\mu_0, \mu_1, c^*, \alpha, r)$ and initial state y_0 as used above. For simplicity, M–noise was ignored in the present study.

To optimize for speed, functions for simulating antibody levels from the longitudinal model, random parameter selection, noise sampling, and calculations of the three statistics for comparing observed and simulated distributions were all implemented in C. A function was also added to translate the D_{KS} into a probability level using an approximation to the Kolmogorov distribution[23,27]. Using this code, a typical evaluation involving $N_{\text{obs}} = N_{\text{sim}} = 2000$ to calculate D_{KS} took 8.9 ms (Linux 5.15; gcc 11.1.0;

Intel Core i7-8559U at 4 GHz).

Approximate Bayesian Computation

As the kinetic parameters were obtained from the separately fitted longitudinal model, there remain three parameters to be estimated: the infection (seroconversion) rate λ and the two parameters (μ, σ) defining the lognormal B–noise distribution.

It is desirable to assess the uncertainty in the parameter estimates jointly. Initial tests showed that any of the statistics used above for evaluating the similarity between observed and simulated samples can be used (see Results below). Of the three metrics, the KS statistic D_{KS} is simplest, both conceptually and computationally. As neither the Anderson–Darling statistic nor the Kullback–Leibler divergence provided clearly superior results only D_{KS} and its associated probability were used.

In the absence of a likelihood function, Approximate Bayesian Computation (ABC) can be used to obtain posterior estimates. Instead of a simple rejection algorithm[28], we have employed a Metropolis–Hastings sampler to improve efficiency and reduce the computational burden[29]. D_{KS} can be directly translated into a probability p_{KS}, which may be plugged into the Metropolis sampler.

By defining a parameter vector $\boldsymbol{\theta} = (\log(\lambda), \mu, \log(\sigma))$, the product of the prior density $\phi(\boldsymbol{\theta})$ and the Kolmogorov probability

$$\phi(\boldsymbol{\theta}) p_{\text{KS}}(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{sim}}(\boldsymbol{\theta}))$$

determines the ratio of "posteriors" between successive iterations[30]. Here $\boldsymbol{Y}_{\text{obs}}$ is the observed antibody sample and $\boldsymbol{Y}_{\text{sim}}(\boldsymbol{\theta})$ is a simulated sample of antibody concentrations.

Typically, simulations used a burn-in of 1,000 iterations, followed by 2,000 iterations for obtaining a posterior sample. It was easily checked that in a converged Markov chain iterated samples return a high Kolmogorov probability (≥ 0.05 for instance), ensuring that the simulated EDF always matches the observed EDF.

Results

All results shown are based on longitudinal data of serum IgG against pertussis toxin (IgG-PT) also used in earlier studies of seroincidence[31]. Minor adaptations in the longitudinal model[16] to improve parameter estimation are documented in the Appendix section A.1.

Simulated data: exploring metrics

Figure 2a,c show the densities of simulated cross-sectional antibody levels for 2000 subjects sampled at 0–80 years of age (uniformly distributed), using an infection rate $\lambda_{\text{sim}} = 0.001$ (1/yr) and 0.05 (1/yr) respectively. These are treated as the observed data, and compared with simulated data (2000 subjects, same ages) for a range of incidences λ. The fitted densities corresponding to the minimum D_{KS} and D_{KL} are shown in Figures 2a,c. AD deviates are close to the KL divergence estimates. Figure 2b,d shows the KS and AD deviates and the KL divergence as a function of λ: the minima all coincide. In order to show the three metrics on a scale from 0 to 1 they are scaled: for D_{KS}

$$\text{KS Dev} = \frac{D_{\text{KS}}(\lambda) - \min(D_{\text{KS}})}{\max(D_{\text{KS}}) - \min(D_{\text{KS}})}, \tag{12}$$

and the same for the AD deviance A^2 and KL divergence D_{KL}. The probability p_{KS} associated with D_{KS} is also shown as the red line in Figures 2b,d. At lower infection frequency the noise dominates the antibody distribution and estimation of λ shows greater uncertainty (Figure 2a,b) than with higher λ_{sim} (Figures 2c,d).

Figures 2 and 3 here.

Figure 3 shows statistics as a function of the B–noise parameters, at a given λ_{sim}. Figures 3a and 3c for the lognormal (log) mean μ, keeping the log sd at its optimum $\sigma = 0.6$. In the Figures 3b and 3d for the lognormal (log) sd σ, keeping the log mean at its optimum $\mu = 0.1$. When $\lambda_{\text{sim}} = 0.001$ (1/yr) there are few infections and most antibody levels in the cross–sectional sample originate from the B–noise distribution. At $\lambda_{\text{sim}} = 0.05$ (1/yr) B–noise constitutes a smaller fraction of the cross–sectional antibody levels and estimation is more uncertain (compare Figures 2b and 2d).

Simulated data: uncertainty in joint parameter estimates

Figure 4a shows estimated λ_{est} for simulated cross–sectional data ($N_{\text{obs}} = 1000$) at a range of incidences λ_{sim}, accounting for elevated baseline $y_0 = y(t_{\text{inf}})$ due to previous infections. Estimation here involved ABC using univariate wide priors ($N(\log(0.1), 5.0)$) for $\log(\lambda)$ and $N(0.1, 0.5)$ for μ and $N(\log(0.6), 0.5)$ for $\log(\sigma)$) to obtain posterior estimates of $\boldsymbol{\theta} = (\lambda, \mu, \sigma)$. The graphs in Figures 4b and 4c show estimates of the corresponding B–noise parameters (μ, σ). All graphs in Figure 4 show posterior predictive means and 95% quantile ranges. Figure 4a also shows estimates of λ obtained by using the published seroincidence method[14](ML) on the same simulated cross–sectional data. Because the new method based on adjusting simulated distributions uses the shape of the whole EDF of the simulated sample, including the B–noise component, estimates of the parameters of the noise distribution can be used to adjust the B–noise parameter in the seroincidence likelihood[14].

When λ_{sim} is small (< 0.01 1/yr) there are few infections, and the uncertainty in λ_{est} increases, particularly in the KS estimates, compared to the likelihood method. At low incidences the B–noise dominates the cross–sectional data and as λ_{est} becomes more uncertain, estimates of noise parameters become more precise. Conversely, at higher incidences antibody noise represents a minor contribution to the cross-sectional antibody data and estimation of noise parameters is more uncertain.

Figure 4 here.

At high incidences an increasing number of infections occur before antibody concentrations have returned to low baseline levels. The resulting "small jump" seroconversion tends to decrease cross–sectional antibody concentrations and the likelihood method underestimates λ (Figure 4a).

Figure 5 here.

The likelihood method for estimating seroincidence ignored remaining $y(t_{\text{inf}})$ at time of infection. When using the EDF method with a seroresponse model that ignores baseline levels from previous episodes ($y(t_{\text{inf}}) = y_0$) to fit a cross-sectional sample generated with the same model including baseline $y(t_{\text{inf}})$ from past infections, both methods exhibit similar bias (Figure 5a). Conversely, when the simulated cross–sectional sample is generated with a model with fixed baseline y_0 the EDF method based on minimizing p_{KS} shows positive bias, while the likelihood method produces unbiased estimates, even at very high infection frequencies λ_{sim} (Figure 5b).

At low incidences ($\lambda < 0.01$ 1/yr) the likelihood estimator seems to show positive bias. Such bias may be due to imperfect characterization of (B–) noise. At these low incidences there are few observations resulting from past seroconversion and estimation is sensitive to mis-attribution.

Simulated data: infection history

A homogenous Poisson process assumes the seroincidence λ is fixed. Variation in λ among subpopulations, including spatial location, has been studied[21,32,13]. It is likely that during the lifetime of a subject the infection pressure changes, either periodically due to seasonal factors[12] or transiently, for instance during an outbreak. For vaccine preventable diseases many individuals in an exposed population may be vaccinated at some time, also causing seroconversion.

Seroconversion events may be linked to age, as in vaccination schedules, or they may occur at a given calendar date.

To study the influence of seroconversion at a given age on estimates of λ, a vaccine induced seroconversion was simulated, identical to a natural infection, at a fixed age. In addition to their natural infections occurring with rate λ all simulated subjects were assumed to have a vaccine induced seroconversion at age 2 (years), assuming universal compliance (100% uptake).

Figure 6a–c shows estimates λ_{est} and the noise parameters $(\mu_{\text{est}}, \sigma_{\text{est}})$ for a vaccinated population, for a range of simulated infection rates λ_{sim} like in Figure 4. Figure 6d–f shows the same estimates when the fitted model ignores vaccination. Obviously, at low infection rates the vaccination event causes substantial overestimation of λ, similar to that for the likelihood based estimate. When $\lambda_{\text{sim}} \gtrapprox 0.5$ (1/yr) the vaccine induced elevation of baseline y_0 is drowned out by later seroconversions and the bias in λ_{est} decreases.

During an outbreak infection rates may rapidly increase, reach a peak level and than decline again, reflecting the epidemic curve[17]. Infection intervals corresponding to a transient peak in λ may be simulated as a sample from a non–homogeneous Poisson process[33,34], see Appendix Figure B7. A recent outbreak causes estimates of λ to be considerably higher, during a period determined by antibody decay (Appendix Figure B8). This increase also depends on the ages of subjects. Obviously, when an individual was born after the outbreak occurred, their antibodies remain at post–outbreak baseline levels, compare Figures B8a and e. The mismatch between simulated and fitted models (that assume a homogeneous Poisson process) can be detected by p_{KS}, indicating failure to achieve a close fit to the "outbreak" data (Figures B8d and h). Such goodness of fit indicators, combined with possible epidemiological information, provides ample opportunity to indicate possible bias caused by past outbreaks.

When the infection pressure is changed due to human intervention, as for instance quarantine or hygiene measures, the decrease in λ may be estimated in a cohort study setting. Suppose there are two study populations, both exposed to baseline infection pressure resulting in infection rate λ_{pre}. At some time, say T years before sampling, one of the study groups then is treated by an intervention causing a drop in infection rate to λ_{post}. Then both groups are sampled and their serum antibodies measured. Both the λ_{pre} and λ_{post} may now be estimated. Two scenarios were compared: a high risk setting with $\lambda_{\text{pre}} = 1.0$ (1/yr) decreased to $\lambda_{\text{post}} = 0.4$ (1/yr) and a lower risk setting with $\lambda_{\text{pre}} = 0.1$ (1/yr) decreased to $\lambda_{\text{post}} = 0.01$ (1/yr). Table 1 shows how age is important in this design: for subjects aged 0 – 10 years both λ_{pre} and λ_{post} can be estimated reasonably well (judged by 95% posterior intervals). In adults (aged 20 – 80 years) only λ_{pre} can be estimated while the post–intervention infection rate λ_{post} is highly uncertain.

Observed data: ESEN

Data from a population study on pertussis in EU countries using standardized units for IgG–PT concentrations in serum[35,36] can be used to examine the performance of the simulation–based estimation method in a practical context. Serum antibody population data for the Germany, Finland, France, Italy, the Netherlands and the United Kingdom were collected between 1994 and 1998, study details are available[35].

To illustrate the fitted distribution samples, Figure 7 shows densities and the test statistics for the Dutch sample, ages 35–40 yr ($N_{\text{obs}} = 502$). The simulated cross–sectional antibody levels match closely with the observed distributions at all ages ($327 \leq N_{\text{obs}} \leq 923$; see Figures 7b,B3 and B6a).

Figure 7 here.

Using the EDF method and parameter estimation with ABC, λ and the B–noise parameters may be estimated, as shown in Figures 8 and 9. Although the estimated seroconversion rate by EDF (Figures 8a and 9a) is slightly higher than the estimate from the likelihood method, both are somewhat lower than earlier estimates[36], where the contribution of B–noise was ignored. Due to the low infection rate, $\lambda_{\text{est}} \approx 0.01$ (1/yr), there are few seroconversions in the youngest age category (0–5 yrs) and estimates of λ are uncertain, as is apparent in Figures 8a and 9a.

Figures 8 and 9 here.

In Figures 8a and 9a estimates of λ by means of the EDF based ABC method are shown together with maximum likelihood estimates using the method described in[14] with noise parameter $\nu = 3.0$ (IU/ml) which corresponds to the 0.95 quantile of a lognormal distribution with parameters $(0.1, 0.6)$. When the noise parameters are adjusted to the estimates obtained with the EDF method for each age category, the likelihood estimates of λ_{est} are (slightly) shifted towards those obtained with the EDF method.

The age patterns in seroconversion rates estimated by ABC (Figures 8a and 9a) do not differ much from the likelihood estimates. Antibody (B) noise shows some variation, within countries by age, but also between countries.

To check how vaccination might have interfered with these estimates these observed data were also fitted by the model that included universal vaccination at age 2 (years). Results are shown in the Appendix, Figures B4 – B5. As expected, the estimated infection rates are lower, in particular in younger subjects. It should be noted, however, that for these vaccination model simulations the posterior Kolmogorov probabilities were very low, except in subjects older than 50 years. Simulations ignoring vaccination appear to better fit these cross–sectional population data , see Figure B6a,b for an example. Such a mismatch indicates that this vaccination model is a poor fit to the observed data, possibly because of a lack of vaccine induced IgG–PT seroresponses[37].

Discussion

The present paper enhances and refines methods for estimating seroconversion rates in a population exposed to an infectious pathogen. Accounting for residual antibody levels from previous infections allows for adaptation at high infection rates, potentially removing bias. Fitting by means of comparing empirical distribution functions provides goodness–of–fit information, valuable in model selection. And finally, the simulation approach allows for arbitrary infection patterns, providing a basis for comparing different infection histories using population serology data.

Employing a longitudinal model to forecast expected antibody levels in cross–sectional population studies requires making substantial assumptions about the seroresponses that generate the observed antibody levels within the study population[10]. So far, these two–stage methods for seroincidence estimation have not dealt with the question how well the distribution of antibody levels in cross–sectional population samples is approximated by the predictions based on the kinetic seroresponse model[5,14]. Here it is demonstrated that the empirical distribution functions of simulated samples of cross–sectional antibody concentrations can be made to closely match observed distributions, by adjusting only the infection (Poisson) rate parameter λ and the parameters of the B–noise distribution. Remarkably, in posterior MC samples the Kolmogorov $\hat{p}_{\mathrm{KS}}$ usually is higher than 0.5, indicating close agreement between observed and simulated antibody distributions (See Appendix Figure B6a). Thus, the longitudinal model accurately predicts the marginal distribution of antibody levels in a cross–sectional population sample.

A previous study underscored the significant influence of B–noise on results, a finding that aligns with the understanding that fitting involves comparing the full spectrum of serum antibody levels, including low levels typically classified as sero–negative[1]. For pragmatic reasons the characterization of B–noise in the likelihood method assumed a uniform distribution[14]. The current approach demonstrates how, at low incidences, estimates of λ are somewhat sensitive to adjustment of the B–noise parameters. M–noise was ignored here but it would not be difficult to include it into the simulations. Even though estimating its magnitude, or even the shape of its distribution, can be expected to be difficult, repeated measurements of control sera in the laboratory should allow quantitative specification of the M–noise associated with the used assay.

The simulation approach allows greater freedom in assumptions about the infection history of individual subjects in the population sample. For pertussis, including the infection history of a subject to determine the antibody level at the start of the sampled infection episode appears to not strongly influence the estimated seroconversion rates, as was noted earlier[14], because of the (relatively) low infection pressure of pertussis in the population. However, Figure 2 shows that at higher infection pressures elevated baseline concentrations can cause considerable bias leading to underestimation of the sero–incidence. Such high infection pressures may occur in other pathogens, e.g., *Campylobacter*[21] or typhoid *Salmonella*[19], or transiently during outbreaks[38,39].

It should be noted that mortality may also lead to bias in estimates of λ: individuals who die will not be captured in cross–sectional samples leading to underestimation. When mortality is low this effect may be ignored.

A basic assumption for the likelihood approach to seroincidence was a population in endemic equilibrium: in the exposed population infections were assumed to occur as a stationary Poisson process[5,14]. Seasonality, if present in infection pressure, has previously been accounted for[12]. The observed variation by age in seroconversion rates can be explained by variation in contact patterns that drive transmission[36]. Here we also find variation in baseline noise, possibly due to differences in infection history among the study populations. When cross–sectional data include multiple antibodies for each blood sample it is often possible to estimate λ in very small sample sizes[13]. In population samples for *Campylobacter* and nontyphoid *Salmonella* the variation in λ among individual subjects appeared small, not inconsistent with Poisson incidence. Nevertheless, there are many settings with strong variation in infection pressure, not in the least during outbreaks of infectious disease[39]. Seroconversions leading to elevated serum antibody levels at the time of infection may also result from vaccination[37], interfering with the estimates of λ in young children.

Earlier likelihood based analyses included only the most recent infection episode, therefore any variations in λ before the most recent infection could not be accounted for. The simulation approach does not have this limitation. The examples included here show how earlier seroconversions may influence estimates of λ, depending on age of sampled subjects and serum antibody decay rates. For instance in pertussis, when infection rates are higher than approximately 0.1 (1/yr), universal vaccination at an early age does not cause important bias in estimates of λ. Of course, a recent outbreak would affect λ and estimates may represent a weighted average over the time course of λ[40]. However, the age distribution of the exposed population and their individually varying seroresponses interfere with estimates of λ. As outbreaks are often detected, cases may be sampled during the event, so that the sampling dates are known relative to the timing of the outbreak, and a backcalculation approach may be a more informative alternative[17].

It seems possible that there are many spatiotemporal patterns for λ that lead to matching posterior marginal distributions for an observed cross–sectional serum antibody sample. Transient peaks in λ as during outbreaks could be driven by a transmission model, generating individual sequences of infection events accounting for person–to–person transmission. Other events that influence λ, like for instance vaccination schedules (but also vaccine uptake and their composition), or timing of interventions. The examples given above show how gross changes in infection rate may influence estimation of λ. Similar scenarios may be adjusted to specific settings. The goodness of fit check that is implicit in the simulation procedure then becomes even more helpful, providing criteria for accepting or rejecting models.

The function $y_1 = f(y_0)$ (Figure A3a) quantifies the influence of the baseline antibody concentration y_0 at the time of infection on the peak level of the subsequent seroconversion[15]. This can be interpreted as a model for immune "memory": repeated seroconversions start from an elevated level but remain relatively small in magnitude while low y_0 after a long interval between infections leads to a strong response to high peak levels. Cross–

sectional antibody samples generated by this model appear to fit observed population samples quite well but one would like to have observational evidence for the correct shape of the function $f(y_0)$. Similarly, antibody decay may be different in secondary versus primary infections. Different decay rates would strongly influence seroincidence estimates[16]. Establishing this would require observations including repeated infections in any single subject with the same pathogen. As Figures 8 and 9 show this is not likely for pertussis. A sensitivity analysis using age–dependent decay rates might provide insights on how such dependence might affect estimates of λ.

Aside from removing bias, accounting for the infection history in seroincidence estimation allows for distinguishing "small" and "large" seroconversions, possibly associated with subclinical and clinical infections. Given a certain duration of acute symptoms, one could calculate numbers of illnesses in a population. Thus, an attack rate (of symptomatic infections) can be related to the frequency of seroconversion, that is: the infection frequency including asymptomatic infections. When acute symptomatic cases in the observed population are known, estimates of the fraction symptomatic infections could be used as an indirect indication of the validity of the seroresponse model for repeated infections.

As subjects age, their likelihood of having had more than one infection increases, thereby causing age–dependent seroresponses, but it is also possible that the kinetics of seroresponses vary naturally with age due to the development of the immune system[19]. The kinetic parameters $(\mu_0, \mu_1, c^*, \alpha, r)$ may then vary with age of a subject, or more specifically, the infection history of a subject may change any of the kinetic parameters. A limitation of this model is that it ignores the boosting effect of secondary exposures. During the infection phase the rate of antibody production is expected to be higher with secondary exposures and consequently the rate of pathogen growth may be slower[41]. Similarly, the numbers of antibody production sites may change upon repeated infections thus modifying the slope of antibody decay[16]. With such age–dependent responses it would be hard to obtain a marginal model for distributions of antibody levels in a cross–sectional population sample[12]. The simulation approach introduced here allows for the construction of individual trajectories of serum antibody levels during the entire history of any individual in the chosen sample. Therefore, there is no impediment to a model with age dependent seroresponse parameters, except for a performance hit due to the required additional computations.

Funding

This work was partially supported by the Fogarty International Center at the National Institutes of Health [K01 TW012177-01A1] to KA.

Data and code availability

Longitudinal data on pertussis seroresponses are property of RIVM and have been made publicly available on request and can be obtained from the first author.

The ESEN pertussis data are property of the Public Health Institutions of the contributing countries[35].

Pending a new "serocalculator" package on CRAN all code for the simulations is publicly available on Codeberg.

The longitudinal model

`https://codeberg.org/peter19/pertussis/longitudinal/`

Simulation of cross–sectional samples and ABC sampling

`https://codeberg.org/peter19/serocalc/`

Additional scripts to run the pertussis simulations

`https://codeberg.org/peter19/pertussis/serocalc/`

References

1. de Greeff SC, Teunis P, de Melker HE, et al. Two-component cluster analysis of a large serodiagnostic database for specificity of increases of IgG antibodies against pertussis toxin in paired serum samples and of absolute values in single serum samples. Clinical and Vaccine Immunology 2012;19(9):1452–1456. doi:10.1128/CVI.00229-12.

2. Konda T, Kamachi K, Iwaki M, Matsunaga Y. Distribution of pertussis antibodies among different age groups in Japan. Vaccine 2002;20:1711–1717.

3. Nardone A, Pebody RG, Maple PAC, Andrews N, Gay NJ, Miller E. Sero-epidemiology of *Bordetella pertussis* infections in England and Wales. Vaccine 2004;22(9-10):1314–1319. doi: 10.1016/j.vaccine.2003.08.039.

4. Peasey AE, Ruiz-Palacios GM, Quigley M, et al. Seroepidemiology and risk factors for sporadic norovirus/Mexico strain. Journal of Infectious Diseases 2004;189(11):2027–2036.

5. Teunis PFM, van Eijkeren JCH, Ang CW, et al. Biomarker dynamics: estimating infection rates from serological data. Statistics in Medicine 2012;31(20):2240–2248. doi:10.1002/sim.5322.

6. Glynn MK, Friedman CR, Gold BD, et al. Seroincidence of *Helicobacter pylori* infection in a cohort of rural Bolivian children: acquisition and analysis of possible risk factors. Clinical Infectious Diseases 2002; 35(9):1059–1065. doi:10.1086/342910.

7. Hagan H, Thiede H, Des Jarlais DC. Hepatitis C virus infection among injection drug users: survival analysis of time to seroconversion. Epidemiology 2004;15(5):543–549.

8. Farrington CP. Modelling forces of infection for measles, mumps and rubella. Statistics in Medicine 1990;9(8):953–967.

9. Shkedy Z, Aerts M, Molenberghs G, Beutels P, van Damme P. Modelling age–dependent force of infection from prevalence data using fractional polynomials. Statistics in Medicine 2006;25(9):1577–1599. doi: 10.1002/sim.2291.

10. Arnold BF, van der Laan MJ, Hubbard AE, et al. Measuring changes in transmission of neglected tropical diseases, malaria, and enteric pathogens from quantitative antibody levels. PLoS Neglected Tropical Diseases 2017;11(5):e0005616. doi:10.1371/journal.pntd.0005616.

11. Longini IM, Byers RH, Hessol NA, Tan WY. Estimating the stage-specific numbers of HIV infection using a Markov model and back–calculation. Statistics in Medicine 1992;11(6):831–843. doi: 10.1002/sim.4780110612.

12. Simonsen J, Mølbak K, Falkenhorst G, Krogfelt KA, Linneberg A, Teunis PF. Estimation of incidences of infectious diseases based on antibody measurements. Statistics in Medicine 2009;28(14):1882–1895. doi:10.1002/sim.3592.

13. Monge S, Teunis P, Friesema I, et al. Immune response-eliciting exposure to *Campylobacter* vastly exceeds the incidence of clinically overt campylobacteriosis but is associated with similar risk factors: A nationwide serosurvey in the Netherlands. The Journal of Infection 2018; 77(3):171–177. doi:10.1016/j.jinf.2018.04.016.

14. Teunis PFM, van Eijkeren JCH. Estimation of seroconversion rates for infectious diseases: Effects of age and noise. Statistics in Medicine 2020;39(21):2799–2814. doi:10.1002/sim.8578.

15. de Graaf WF, Kretzschmar MEE, Teunis PFM, Diekmann O. A two-phase within-host model for immune response and its application to serological profiles of pertussis. Epidemics 2014;9:1–7. doi: 10.1016/j.epidem.2014.08.002.

16. Teunis PFM, van Eijkeren JCH, de Graaf WF, Bonačić Marinović A, Kretzschmar MEE. Linking the seroresponse to infection to within-host heterogeneity in antibody production. Epidemics 2016;16:33–39. doi: 10.1016/j.epidem.2016.04.001.

17. Wielders CCH, Teunis PFM, Hermans MHA, van der Hoek W, Schneeberger PM. Kinetics of antibody response to *Coxiella burnetii* infection (Q fever): Estimation of the seroresponse onset from antibody levels. Epidemics 2015;13:37–43. doi:10.1016/j.epidem.2015.07.001.

18. Vaitkeviciute I, Teunis PFM, van Pelt W, Krogfelt KA. Kinetics of serum antibodies in response to infection with *Yersinia enterocolitica*. Epidemiology and Infection 2019;147(e165):1–7. doi: 10.1017/S0950268819000530.

19. Aiemjoy K, Seidman JC, Saha S, et al. Estimating typhoid incidence from community-based serosurveys: A multicohort study. Lancet Microbe 2022;:1–10doi:10.1016/S2666-5247(22)00114-8.

20. Diekmann O, de Graaf WF, Kretzschmar MEE, Teunis PFM. Waning and boosting: on the dynamics of immune status. Journal of Mathematical Biology 2018;119(2):149. doi:10.1007/s00285-018-1239-5.

21. Teunis PFM, Falkenhorst G, Ang CW, et al. *Campylobacter* seroconversion rates in selected countries in the European Union. Epidemiology and Infection 2013;141(10):2051–2057. doi: 10.1017/S0950268812002774.

22. Conover WJ. Practical nonparametric statistics, Wiley. Third ed., 1999; 369–406.

23. Kolmogoroff A. Confidence limits for an unknown distribution function. Annals of Mathematical Statistics 1941;12(4):461–463.

24. Kuiper NH. Tests concerning random points on a circle. Proceedings of the Koninklijke Nederlands Akademie van Wetenschappen, Series A 1960;63:38–47.

25. Stephens MA. EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association 1974;69(347):730–737. doi:10.2307/2286009.

26. Perez-Cruz F. Kullback–Leibler divergence estimation of continuous distributions. IEEE International Symposium on Information Theory (ISIT) 2008;:1666–1670.

27. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical Recipes in C, The Art of Scientific Computing. Cambridge University Press, second ed., 1992.

28. Jiang B, Wu TY, Wong WH. Approximate Bayesian Computation with Kullback–Leibler divergence as data discrepancy. Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018;PMLR: Volume 84.

29. Marjoram P, Molitor J, Plagnol V, Tavaré S. Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences of the USA 2003;100(26):15324–15328. doi: 10.1073/pnas.0306899100.

30. Gilks WR, Richardson S, Spiegelhalter DJ, eds. Markov Chain Monte Carlo in practice. London: Chapman and Hall, 1996.

31. Versteegh FGA, Mertens PLJM, de Melker HE, Roord JJ, Schellekens JFP, Teunis PFM. Age-specific long-term course of IgG antibodies to pertussis toxin after symptomatic infection with *Bordetella pertussis*. Epidemiology and Infection 2005;133:737–748.

32. Simonsen J, Teunis P, van Pelt W, et al. Usefulness of seroconversion rates for comparing infection pressures between countries. Epidemiology and Infection 2011;139(4):636–643. doi: 10.1017/S0950268810000750.

33. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2023. URL `http://www.R-project.org/`.

34. Brock K, Slade D. poisson: Simulating Homogenous & Non-Homogenous Poisson Processes, 2015. URL `https://CRAN.R-project.org/package=poisson`. R package version 1.0.

35. Pebody RG, Gay NJ, Giammanco A, et al. The seroepidemiology of *Bordetella pertussis* infection in Western Europe. Epidemiology and Infection 2005;133(1):159–171.

36. Kretzschmar MEE, Teunis PFM, Pebody RG. Incidence and reproduction numbers of pertussis: Estimates from serological and social contact data in five European countries. PLoS Medicine 2010;7(6):1–10 (e1000291).

37. Berbers GAM, van de Wetering MSE, van Gageldonk PGM, Schellekens JFP, Versteegh FGA, Teunis PFM. A novel method for evaluating natural and vaccine induced serological responses to *Bordetella pertussis* antigens. Vaccine 2013;31(36):3732–3738. doi: 10.1016/j.vaccine.2013.05.073.

38. Baker JM, Nelson KN, Overton E, et al. Quantification of occupational and community risk factors for SARS-CoV-2 seropositivity

among healthcare workers in a large U.S. healthcare system. Annals of Internal Medicine 2021;174(5):649–654. doi:10.7326/M20-7145.

39. Havers FP, Reed C, Lim T, et al. Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020. JAMA Internal medicine 2020;180(12):1576–1586. doi: 10.1001/jamainternmed.2020.4130.

40. Whitaker HJ, Farrington CP. Estimation of infectious disease parameters from serological survey data: the impact of regular epidemics. Statistics in Medicine 2004;23(15):2429–2443.

41. Roitt IM, Brostoff J, Male DK. Immunology. Mosby, 1993.

42. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). Vienna, Austria, 1–10.

Tables

Table 1: Simulated intervention: 1000 subjects sampled 2 years after an intervention that decreased the infection rate from λ_{pre} to λ_{post} (column "set"). In addition, a control group of 1000 subjects was sampled assuming they were only exposed to the baseline infection rate λ_{pre}. The table shows joint estimates of baseline infection rate λ_{pre}, post–intervention infection rate λ_{post}, and noise parameters $(\mu_{\text{noise}}, \sigma_{\text{noise}})$.

age			estimated		
range (yr)		set	mean	95% range	
0 – 10	λ_{pre}	1.00	1.00	0.83 – 1.19	1/yr
	λ_{post}	0.40	0.27	0.10 – 0.53	1/yr
	μ_{noise}	0.10	0.06	-0.28 – 0.45	
	σ_{noise}	0.60	0.61	0.40 – 1.03	
0 – 10	λ_{pre}	0.10	0.11	0.09 – 0.13	1/yr
	λ_{post}	0.01	0.01	0.00 – 0.06	1/yr
	μ_{noise}	0.10	0.11	0.02 – 0.21	
	σ_{noise}	0.60	0.64	0.53 – 0.78	
20 – 80	λ_{pre}	1.00	1.01	0.87 – 1.15	1/yr
	λ_{post}	0.40	0.06	0.00 – 286.8	1/yr
	μ_{noise}	0.10	0.14	-0.42 – 0.66	
	σ_{noise}	0.60	0.71	0.28 – 1.54	
20 – 80	λ_{pre}	0.10	0.09	0.07 – 0.11	1/yr
	λ_{post}	0.01	0.20	0.00 – 8.20	1/yr
	μ_{noise}	0.10	0.37	-0.06 – 0.93	
	σ_{noise}	0.60	0.75	0.47 – 1.23	

Figures

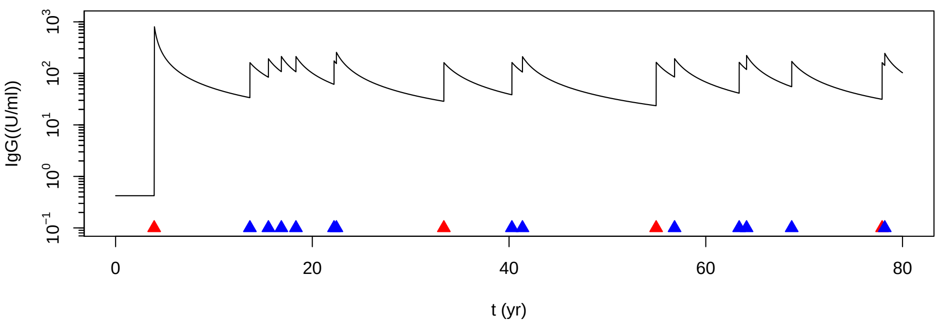
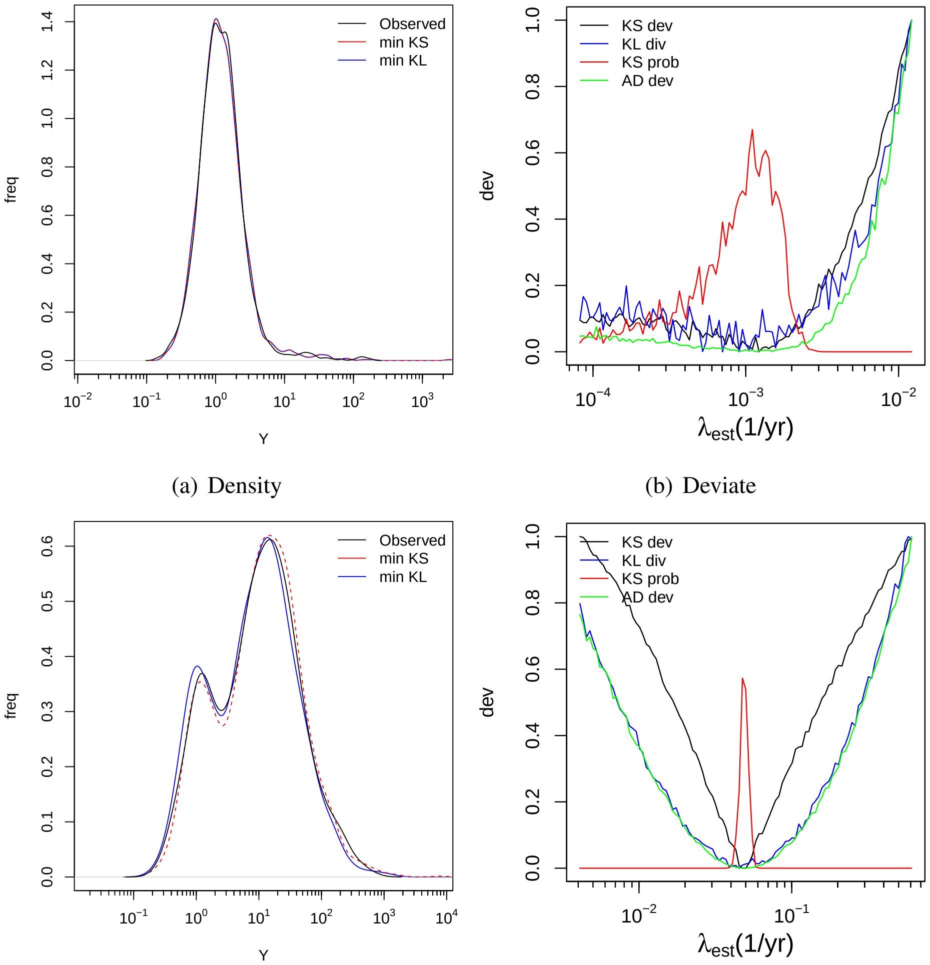


Figure 1: Simulated seroresponse of a hypothetical subject, from birth to age 80 (years), infections occurring as a Poisson process with rate 0.2/yr. Longitudinal parameters fitted to pertussis data[16]: $(\mu_0, \mu_1, c, \alpha, r)$ chosen at birth and kept fixed. The baseline antibody level y_0 is low at birth. After the first infection y_0 is carried over from each prior episode for any further infections. Triangles indicate symptomatic (“large jump”: red) or asymptomatic seroconversions (“small jump”: blue).



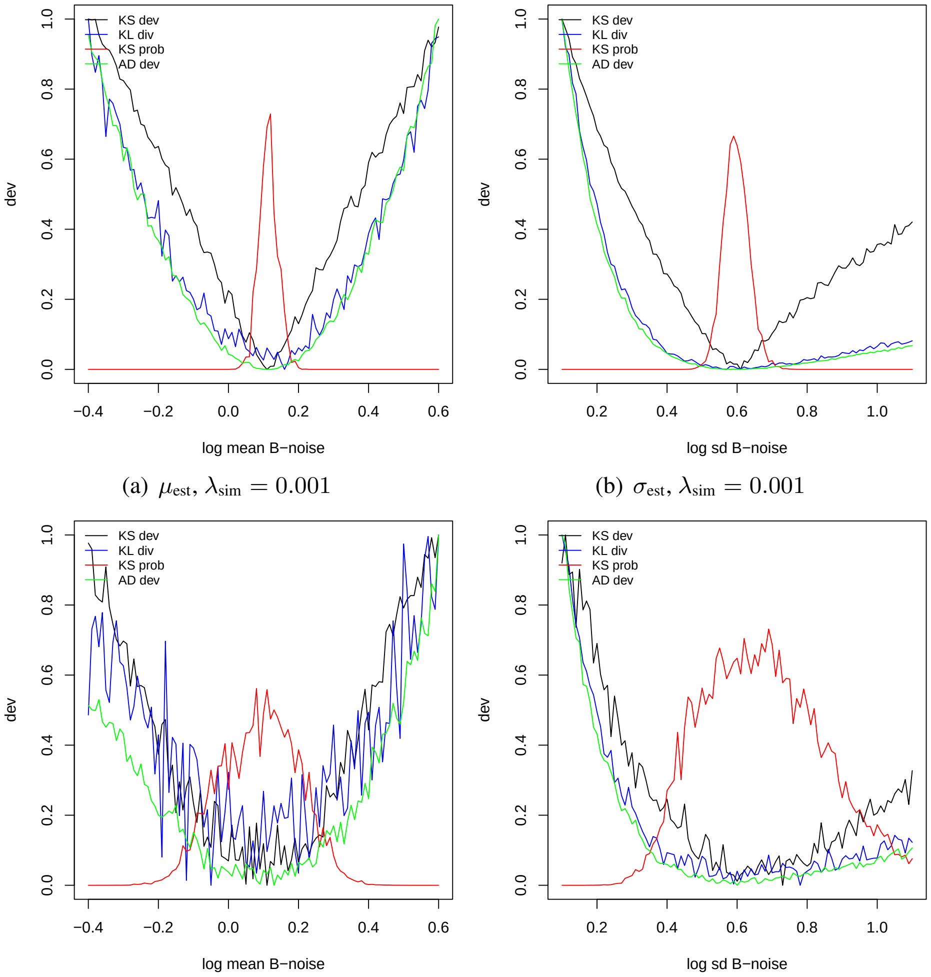
(a) Density

(b) Deviate

(c) Density

(d) Deviate

Figure 2: Fitting λ: output for simulated data for subjects 0 – 80 yrs of age with baseline distribution parameters: (0.1, 0.6) and seroconversion rate λ_{sim} = 0.001 and 0.05 (1/yr). (a) Probability density of (simulated) observed serum antibody distribution and best fitting (minimum D_{KS} or D_{KL}) simulated distributions. (b) Scaled deviates as a function λ_{est} of the simulated sample of antibody concentrations. (c) and (d): corresponding KS dev (Kolmogorov–Smirnov deviate D_{KS}); KL div (Kullback–Leibler divergence D_{KL}); AD dev (Anderson–Darling deviate A^2); KS prob (Kolmogorov probability p_{KS}) as a function of λ.



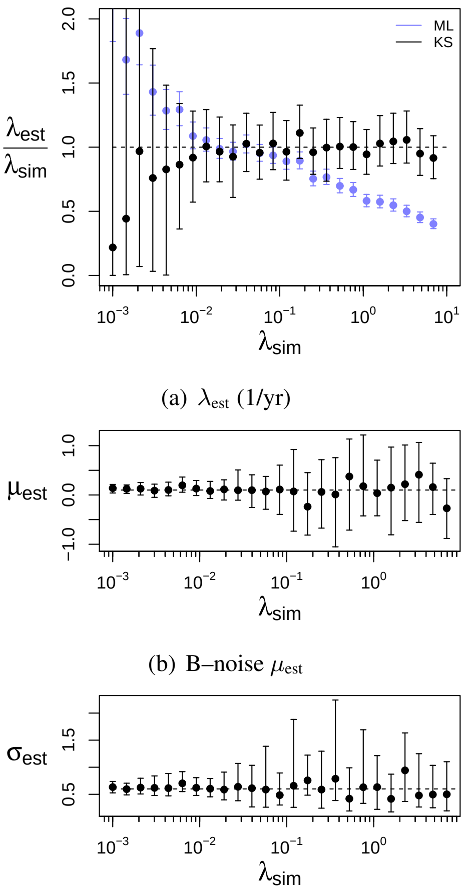
(a) $\mu_{\text{est}}, \lambda_{\text{sim}} = 0.001$

(b) $\sigma_{\text{est}}, \lambda_{\text{sim}} = 0.001$

(c) $\mu_{\text{est}}, \lambda_{\text{sim}} = 0.05$

(d) $\sigma_{\text{est}}, \lambda_{\text{sim}} = 0.05$

Figure 3: Fitting B–noise distribution parameters: output for simulated data for subjects 0 – 80 yrs of age with baseline distribution parameters: (0.1, 0.6) and seroconversion rate λ_{sim} = 0.001 and 0.05 (1/yr). (a) Scaled deviates as a function of the B–noise log mean μ_{est} at $\lambda_{\text{sim}} = 0.001$ (1/yr). (b) Scaled deviates as a function of the B–noise log sd σ_{est} at $\lambda_{\text{sim}} = 0.001$ (1/yr). (c) and (d) Same at $\lambda_{\text{sim}} = 0.05$ (1/yr). KS dev: Kolmogorov–Smirnov deviate D_{KS}; KL div: Kullback–Leibler divergence D_{KL}; AD dev: Anderson–Darling deviate A^2; KS prob: Kolmogorov probability p_{KS}.

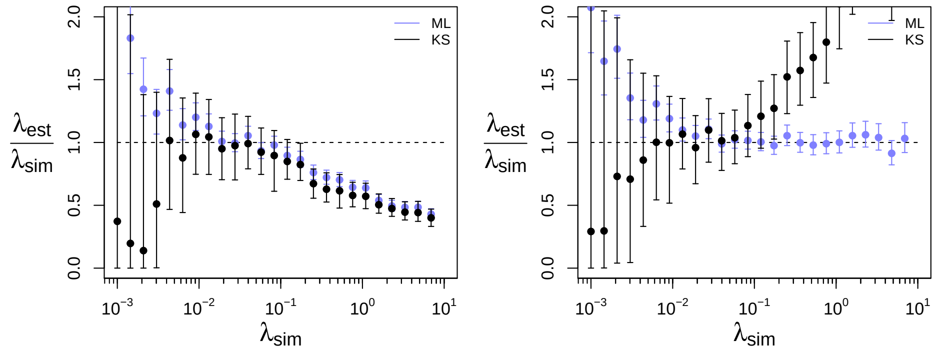


(a) λ_{est} (1/yr)

(b) B–noise μ_{est}

(c) B–noise σ_{est}

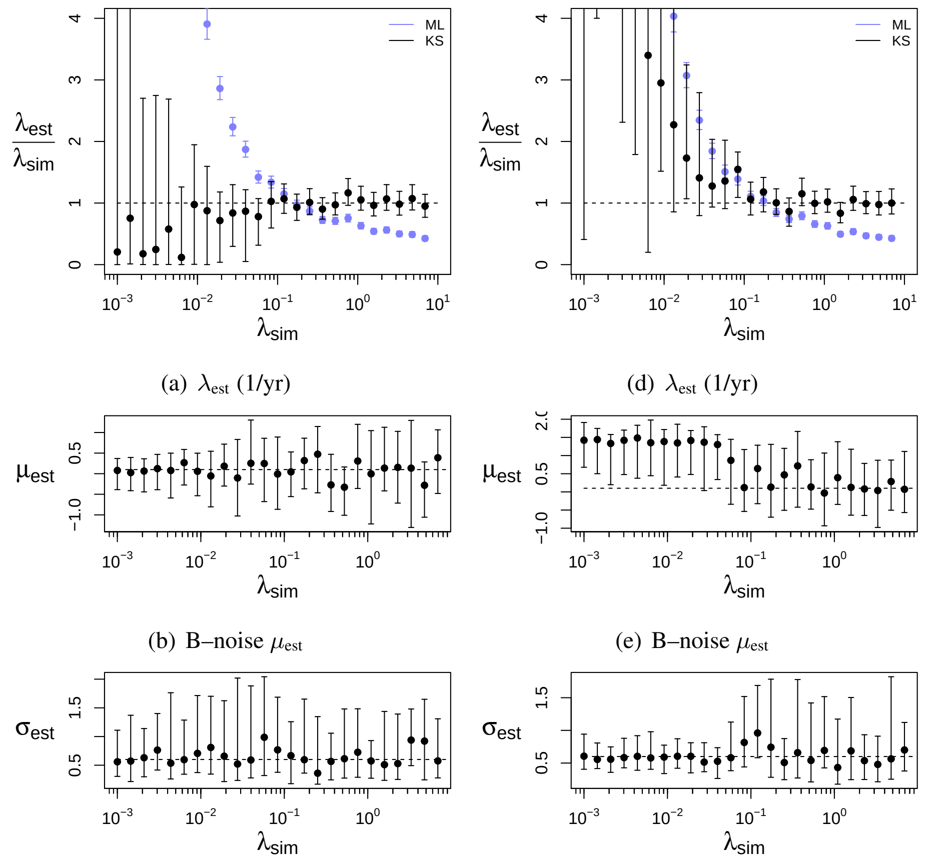
Figure 4: Estimated λ_{est} and B–noise parameters $(\mu_{\text{est}}, \sigma_{\text{est}})$ for a range of simulated λ_{sim} ranging from 0.001 to 10 (1/yr) in subjects 0 – 80 yrs of age and B–noise distribution parameters: $(\mu_{\text{sim}}, \sigma_{\text{sim}}) = (0.1, 0.6)$. Baseline $y_0 = y(t_{\text{inf}})$ from previous infections, generating seroresponses as in Figure 1. (a) Two methods for estimation of λ are compared: ML: maximum likelihood using the published seroincidence method[14] with fixed B–noise parameter adjusted to the 95th percentile of simulated noise ($\nu = 2.97$ IU/ml), and KS: EDF based method using p_{KS} to jointly estimate (λ, μ, σ). (b) B–noise log mean μ_{est} and (c) log sd σ_{est}, estimated jointly with λ_{est} using ABC. Dashed lines indicate simulated values: $\lambda_{\text{est}} = \lambda_{\text{sim}}$, $\mu_{\text{est}} = \mu_{\text{sim}} = 0.1$, $\sigma_{\text{est}} = \sigma_{\text{sim}} = 0.6$.



(a) a

(b) b

Figure 5: Bias in λ_{est} due to baseline "memory". (a) Simulated cross–sectional sample generated with $y_0 = y(t_{\text{inf}})$; λ_{est} calculated by ML and EDF fitting (KS) with fixed baseline ($y_0 = y(0)$) and instantaneous seroconversion ($t_1 = 0$). (b) Simulated cross–sectional sample generated with fixed $y_0 = y(0)$ and instantaneous seroconversion ($t_1 = 0$); λ_{est} calculated by ML and EDF fitting (KS) with infection history ($y_0 = y(t_{\text{inf}})$) and seroconversion ($t_1 > 0$).



(a) λ_{est} (1/yr)

(d) λ_{est} (1/yr)

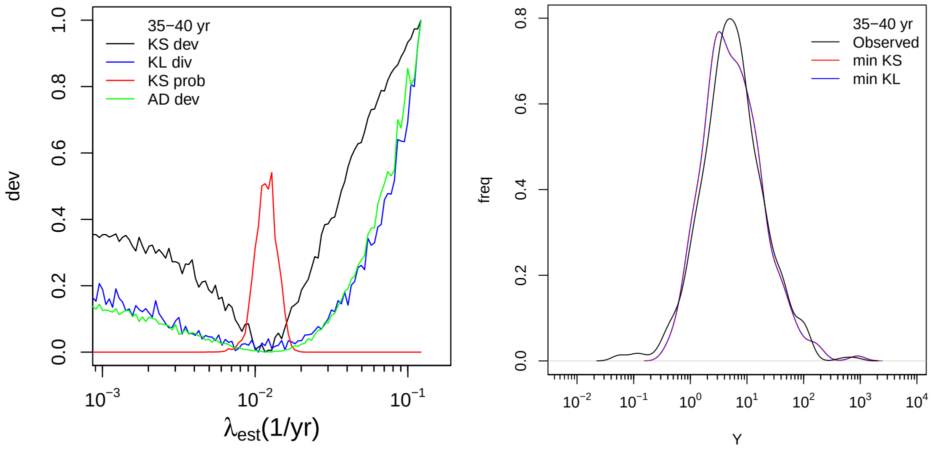
(b) B–noise μ_{est}

(e) B–noise μ_{est}

(c) B–noise σ_{est}

(f) B–noise σ_{est}

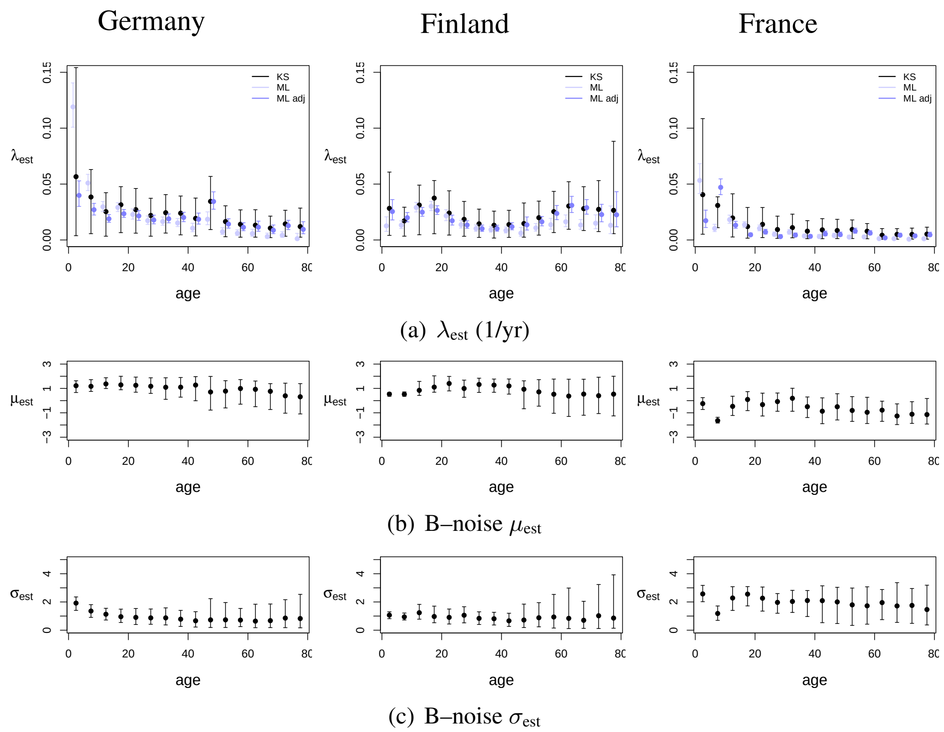
Figure 6: Estimated λ_{est} and B–noise parameters $(\mu_{\text{est}}, \sigma_{\text{est}})$ for a simulated population vaccinated at age 2. a–c: λ_{est} and B–noise parameters $(\mu_{\text{est}}, \sigma_{\text{est}})$ estimated using a model including vaccination (KS). d–f: same parameters estimated using a model without vaccination at age 2 (KS). For comparison, likelihood based estimated are also shown (ML).



(a) λ_{est}

(b) Density of y

Figure 7: ESEN data from the Netherlands: estimates of λ and probability density of antibody levels. Age 35–40 yr (a): D_{KS} (KS dev) and associated probability p_{KS} (KS prob) as a function of λ_{est}. Also shown D_{KL} (KL div) and Anderson–Darling A^2 (AD dev). (b): probability densities of observed antibody levels and minimum D_{KS} (maximum p_{KS}) sample, and minimum D_{KL} sample.

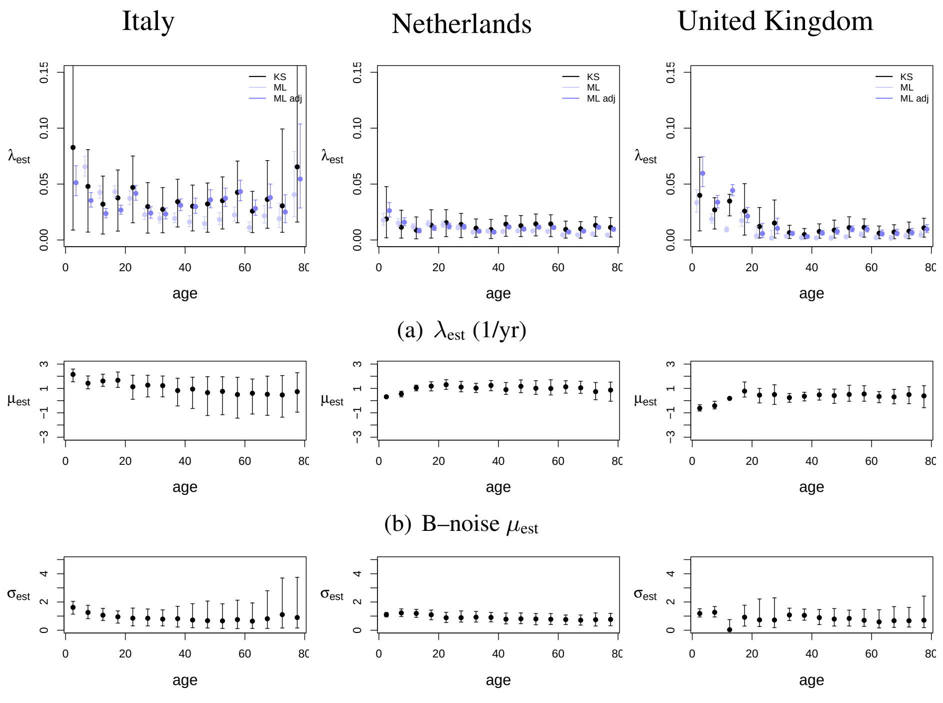


(a) λ_{est} (1/yr)

(b) B–noise μ_{est}

(c) B–noise σ_{est}

Figure 8: ESEN data, (a) Estimated seroconversion rates λ_{est} by (5 yr) age categories. Two fitting methods are shown. ML: maximum likelihood seroincidence[14] with B–noise fixed $\nu = 3.0$ IU/ml. ML adj: maximum likelihood seroincidence with B–noise adjusted to the 95th percentile of the distribution estimated by EDF. KS: EDF method using p_{KS}, fitted by ABC jointly estimating λ and the two noise parameters. (b) Estimated log mean μ of B–noise. (c) Estimated log sd σ of B–noise.



(a) λ_{est} (1/yr)

(b) B–noise μ_{est}

(c) B–noise σ_{est}

Figure 9: ESEN data, (a) Estimated seroconversion rates λ_{est} by (5 yr) age categories. Two fitting methods are shown. ML: maximum likelihood seroincidence[14] with B–noise fixed $\nu = 3.0$ IU/ml. ML adj: maximum likelihood seroincidence with B–noise adjusted to the 95th percentile of the distribution estimated by EDF. KS: EDF method using p_{KS}, fitted by ABC jointly estimating λ and the two noise parameters. (b) Estimated log mean μ of B–noise. (c) Estimated log sd σ of B–noise.

Appendices

A Seroresponse to infection

A.1 Within–host model for the seroresponse

The within host model[16] relates pathogen growth and antibody mediated pathogen inactivation/removal

$$\begin{array}{llll} & \text{Baseline} & \text{Infection} & \text{Decay} \\ \text{pathogens:} & b(0)=b_0 & b'(\tau)=\mu_0 b(\tau)-cy(\tau) & b(\tau)=0 \\ \text{antibodies:} & y(0)=y_0 & y'(\tau)=\mu_1 y(\tau) & y'(\tau)=-\alpha y(\tau)^r. \end{array} \tag{A.1}$$

The baseline antibody concentration is y_0 and the initial (inoculated?) pathogen concentration at time $\tau = 0$ is b_0. Antibody concentrations $y(\tau)$ increase until all pathogens have been removed, at time $\tau = t_1$:

$$t_1 = \frac{1}{\mu_1 - \mu_0} \log\left(1 + \frac{(\mu_1 - \mu_0)b_0}{cy_0}\right). \tag{A.2}$$

From t_1 antibody decay begins, so that the peak antibody concentration y_1 is reached at $\tau = t_1$:

$$y_1 = y_0 \mathrm{e}^{\mu_1 t_1} = y_0 \left(1 + \frac{(\mu_1 - \mu_0)b_0}{cy_0}\right)^{\frac{\mu_1}{\mu_1 - \mu_0}}. \tag{A.3}$$

As the two parameters b_0 and c only appear as the ratio b_0/c a reduced parameter $c^* = c/b_0$ may be defined[15] and

$$\begin{array}{ll} t_1 & = \frac{1}{\mu_1 - \mu_0} \log\left(1 + \frac{\mu_1 - \mu_0}{c^* y_0}\right) \\ y_1 & = y_0 \left(1 + \frac{\mu_1 - \mu_0}{c^* y_0}\right)^{\frac{\mu_1}{\mu_1 - \mu_0}} \end{array}. \tag{A.4}$$

When the model is applied to analyze seroresponse data, the growth rate of virus μ_0 and the antibody efficiency parameter c cannot be observed because the time course of pathogens is not known. The time course of antibody concentrations can be expressed in antibody parameters

$$y(\tau) = \begin{cases} y_0 \mathrm{e}^{\mu_1 \tau} & 0 \leq \tau \leq t_1 \\ y_1 \left(1 + (r-1)\alpha y_1^{r-1}(\tau - t_1)\right)^{-\frac{1}{r-1}} & t_1 \leq \tau \end{cases}. \tag{A.5}$$

Thus, during infection, antibody concentrations increase exponentially from baseline to some peak concentration y_1, at t_1 days post infection. At peak antibody level the pathogens have been removed and decay starts, with a power function. The shape of the decay curve may provide information about the within–host heterogeneity in antibody production[16].

In any individual subject the seroresponse is determined by five parameters that can be estimated based on the observed time course of serum antibody concentrations: the baseline antibody level y_0, the time (from symptom onset) to peak antibody level t_1, the peak antibody level y_1, the decay rate ν, and the shape factor for the decay phase r.

A.2 Reinfection: the role of y_0

Suppose the rate parameters μ_0 (net growth rate of pathogens), μ_1 (rate of increase of serum antibodies), and c ("efficacy" of antibodies in removing pathogens) are known, and valid for any infection (primary or subsequent) of the same host. Then the baseline antibody level at time of infection determines the next peak antibody level[15]

$$y_1 = f(y_0) = y_0 \left(1 + \frac{(\mu_1 - \mu_0)b_0}{c y_0}\right)^{\frac{\mu_1}{\mu_1 - \mu_0}}. \tag{A.6}$$

When the antibody concentration at the time of infection

$$y_0 > y_{\text{min}} = \frac{\mu_0 b_0}{c}, \tag{A.7}$$

the pathogen concentration

$$b(\tau) = b_0 \mathrm{e}^{\mu_0 \tau} - \frac{c y_0}{\mu_1 - \mu_0} \left(\mathrm{e}^{\mu_1 \tau} - \mathrm{e}^{\mu_0 \tau}\right), \tag{A.8}$$

decreases monotonically and the antibody concentration makes a “small jump” (Fig. 3 in[15]) corresponding to a mild (asymptomatic?) infection.

When a subject is infected for the first time, the antibody level at the time of infection is presumed low[14]. This baseline antibody level y_0 together with the infection parameters μ_0, μ_1, c determines the peak level y_1 that is reached following the first infection. The subsequent time course of decay in antibody levels is determined by the decay parameters ν and r. Anytime a person is infected again the antibody concentration at the time of reinfection, the baseline antibody concentration, determines whether this re–infection will lead to a “small” or a “large” jump to the next peak level[15]. As a “small” jump corresponds to more recent prior seroconversion, these two categories of seroconversions have been assumed to represent mild (asymptomatic) or serious (symptomatic) infections. In a population exposed to a given infection pressure (from a specific infectious pathogen), this causes a fraction of that population to be protected from acute illness, due to immunity associated with recent prior infection[20].

A.3 Parameter estimation

The infection phase of the seroresponse is determined by the rate parameters μ_0 and μ_1 and the efficacy parameter c, and the initial conditions y_0 and b_0. The parameters μ_0 and c cannot be directly observed. The (net) rate of antibody increase depends on observables y_0, y_1 and t_1

$$\mu_1 = \frac{\log(y_1) - \log(y_0)}{t_1}. \tag{A.9}$$

As noted earlier[15] the inital pathogen level b_0 only appears in the ratio c/b_0 and therefore we can use the compound parameter

$$c^* = \frac{c}{b_0} \tag{A.10}$$

instead. Observed seroresponses are from infections that are ultimately cleared. It is therefore reasonable to assume that antibodies outcompete pathogens, and $\mu_1 > \mu_0$. For parameter fitting, we assume

$$\mu_0 = \left(\frac{\mathrm{e}^u}{1 + \mathrm{e}^u} \right) \mu_1. \tag{A.11}$$

In longitudinal data the start of infection is usually identified as symptom onset, so that normally all seroresponses should correspond to “large jumps”[15]. This means that for such observed (observable) seroresponses

$$y_0 < y_{\text{min}} = \frac{\mu_0 b_0}{c} = \frac{\mu_0}{c^*}, \tag{A.12}$$

imposing an additional condition on the parameters to be estimated. Thus, for parameter estimation we assume

$$c^* = \left(\frac{\mathrm{e}^v}{1 + \mathrm{e}^v} \right) \frac{\mu_0}{y_0}. \tag{A.13}$$

This leaves 6 variables to be estimated $(y_0, \mu_1, u, v, \alpha, r)$.

A.4 Implementation

The model was specified and run in JAGS[42].

A multivariate normal prior was used for the parameter vector

$(\log(y_0), \log(mu_1), u, v, \log(\alpha), \log(r-1))$,

with wishart prior ($\boldsymbol{\Omega}$) for the precision matrix, as specified below.

```
#            log(y0), log(mu1), v, u, log(alpha), log(shape)-1)
mu.hyp[1,]    <-      c( 0.0,  0.0,  0.0,  0.0, -2.0,  -3.0);  # IgG
prec.hyp[1,,] <- diag(c( 0.05, 0.05, 2.0,  0.05, 0.001, 4.0)); # IgG
omega[1,,]    <- diag(c(10.0, 10.0,  1.0, 10.0, 10.0,   0.2)); # IgG
wishdf[1] <- 20;
```

As in[15] a normal prior with parameters

```
prec.logconc.hyp <- c(4.0,1);
```

was used for the measurement error in serum antibody concentrations. Antibody concentrations below 5 IU/ml were treated as censored (assuming no accurate readout was possible below that concentration).

JAGS source code is given below:

```
model{
  for(subj in 1:nsubj){
    for(test in 1:ntest){
      logy1[subj,test] <- log(y0[subj,test])+
        log(1+(mu1[subj,test]-mu0[subj,test])/(c1[subj,test]*y0[subj,test]))*
        mu1[subj,test]/(mu1[subj,test]-mu0[subj,test])
      t1[subj,test] <- (logy1[subj,test]-log(y0[subj,test]))/mu1[subj,test]
      for(rec in 1:nrec[subj,test]){
        mu.logy[subj,test,rec] <- ifelse(step(t1[subj,test]-trec[subj,test,rec]),
          log(y0[subj,test])+(mu1[subj,test]*trec[subj,test,rec]),
          1/(1-shape[subj,test])*log(exp(logy1[subj,test])^(1-shape[subj,test])-
            (1-shape[subj,test])*alpha[subj,test]*
    (trec[subj,test,rec]-t1[subj,test])))
        logy.cens[subj,test,rec] ~ dinterval(logy[subj,test,rec],cens.lev.log)
        logy[subj,test,rec] ~ dnorm(mu.logy[subj,test,rec],prec.logy[test])
      }
      y0[subj,test]    <- exp(par[subj,test,1])
      mu1[subj,test]   <- exp(par[subj,test,2])
      mu0[subj,test]   <- mu1[subj,test]*
        exp(par[subj,test,3])/(1+exp(par[subj,test,3]))
      c1[subj,test]    <- mu0[subj,test]/y0[subj,test]*
        exp(par[subj,test,4])/(1+exp(par[subj,test,4]))
      alpha[subj,test] <- exp(par[subj,test,5])
      shape[subj,test] <- exp(par[subj,test,6])+1
      par[subj,test,1:ndim] ~ dmnorm(mu.par[test,],prec.par[test,,])
    }
  }
  for(test in 1:ntest){
    mu.par[test,1:ndim] ~ dmnorm(mu.hyp[test,],prec.hyp[test,,])
    prec.par[test,1:ndim,1:ndim] ~ dwish(omega[test,,],wishdf[test])
    prec.logy[test] ~ dgamma(prec.logy.hyp[test,1],prec.logy.hyp[test,2])
  }
}
```

After adapatation for 2×10^5 iterations, The model was run for 10^6 iterations, with thinning 10^3. Four chains were run, producing a final sample of size 4×10^3.

A.5 Longitudinal model results

Note that the longitudinal model used here differs from[15] in two respects: (1) the restrictions on μ_0 and c^* as described above (sec A.3) and (2) power function decay as in[16].

Although the predicted seroresponses in Figure A1 look similar to those reported earlier[16] parameter estimates are slightly different due to the limits on μ_0 and c^*. The predictive parameter samples may be used to calculate predictions for y_1 and t_1 (Figure A2), which match closely those obtained earlier.

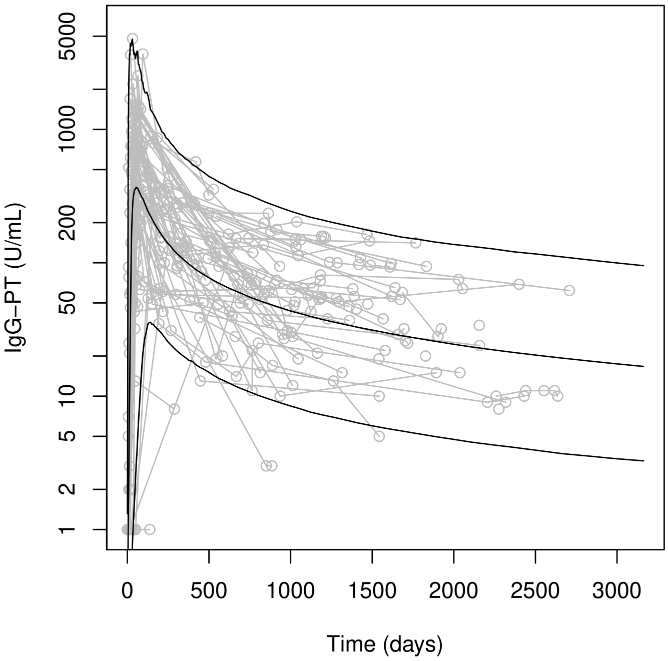
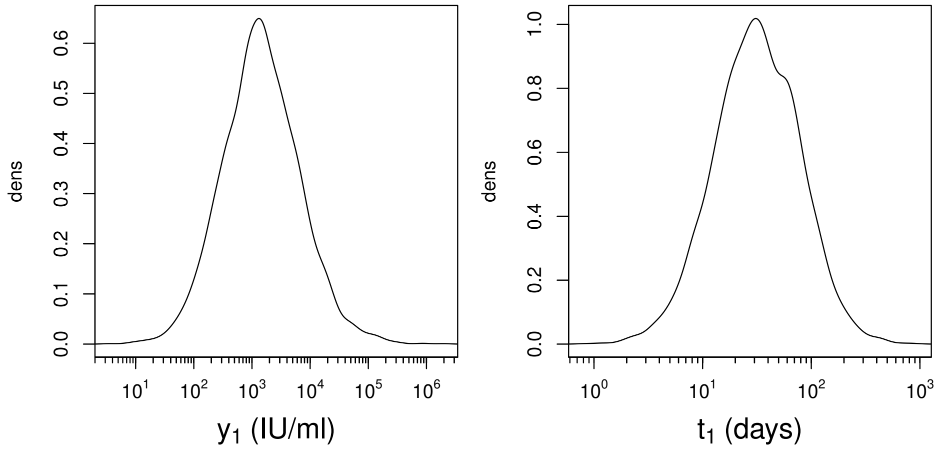


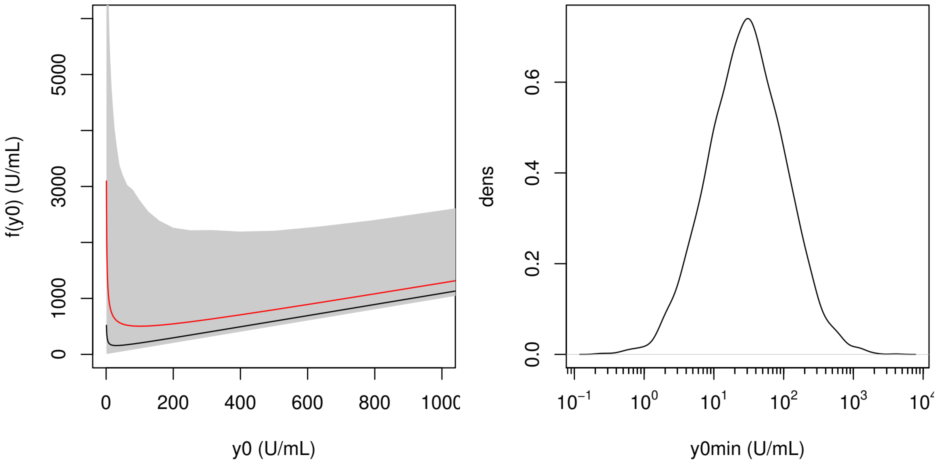
Fig. A1: Predicted seroresponse for the longitudinal model fitted to pertussis antibody data.



(a) y_1 (b) t_1

Fig. A2: Posterior predictive density estimates of the peak antibody level and the time to peak, calculated from the estimated parameters (μ_0, μ_1, c^*) and y_0.

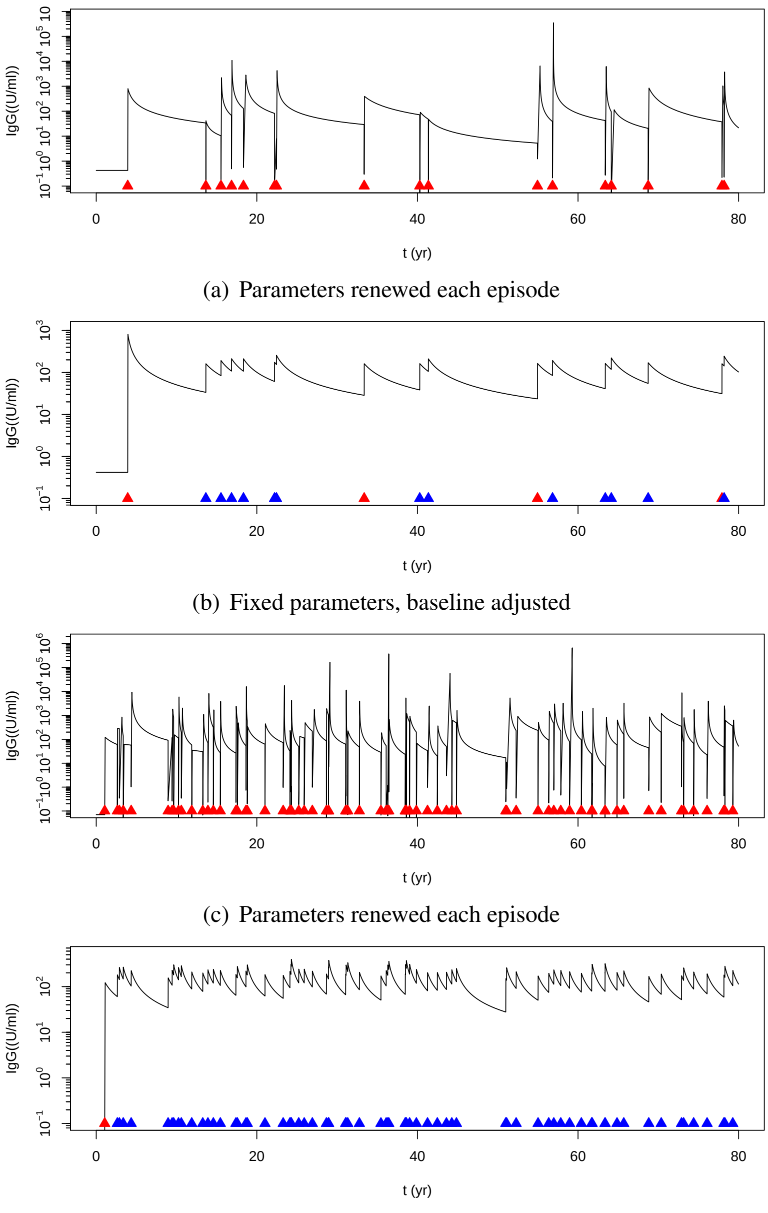
Relevant for the present study is the relation between current baseline y_0 and upcoming peak level y_1 (termed $f(y_0)$ in[15]). This relation is shown in Figure A3 together with predicted levels of $y_{0,\text{min}}$, the threshold above which the next seroconversion is a "small jump'.



(a) $f(y_0)$ (b) $y_{0,\text{min}}$

Fig. A3: (a) Relation between serum antibody baseline y_0 and subsequent peak level $y_1 = f(y_0)$. (b) Distribution of baseline threshold for a subsequent "small jump" in seroresponse.

The estimates obtained with this model ensure a first infection that always is a “large jump”; subsequent infections may cause large or small seroconversions depending on y_0 at the time of infection. Examples are shown in Figure A4. For reference, the alternative model using a new sample from the posterior predictive parameters for each infection episode is also shown, using an identical infection history.



(a) Parameters renewed each episode

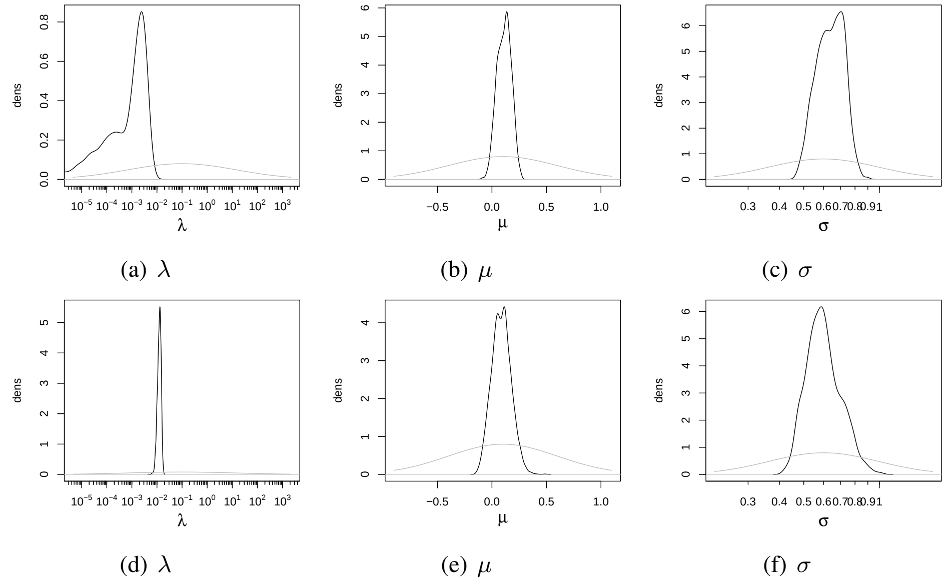
(b) Fixed parameters, baseline adjusted

(c) Parameters renewed each episode

(d) Fixed parameters, baseline adjusted

Fig. A4: Simulated seroresponse of a hypothetical subject, from birth to age 80 (years), infections occurring as a Poisson process with rate 0.2/yr (a,b) and 1/yr (c,d). Longitudinal parameters fitted to pertussis data[16] (a,c): parameters and baseline chosen at random at each new infection. This corresponds with published analyses. (b,d): parameters $(\mu_0, \mu_1, c, \alpha, r)$ and baseline y_0 chosen at birth and kept fixed. Baseline antibody level y_0 for subsequent infections carried over from the prior episode. Triangles indicate symptomatic (“large jump”: red) or asymptomatic seroconversions (“small jump”: blue).

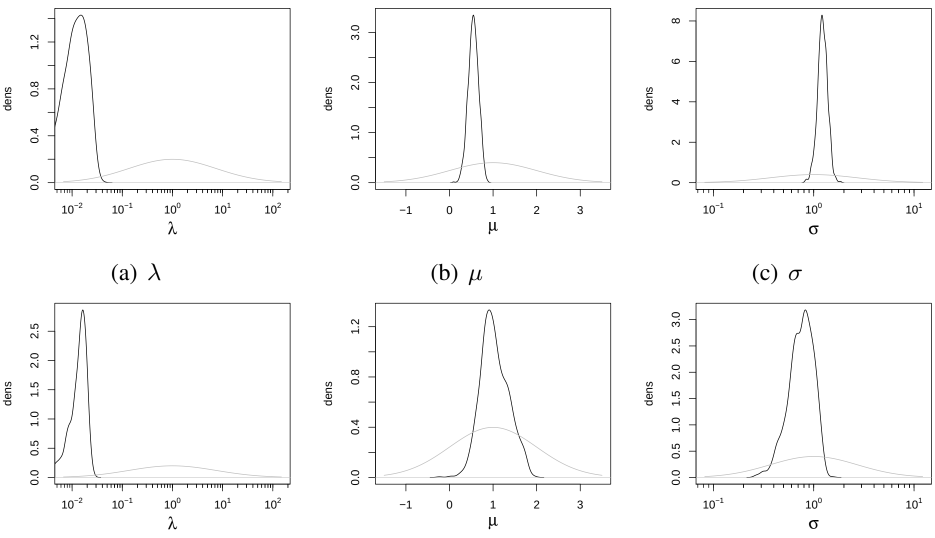
B Additional figures



(a) λ (b) μ (c) σ

(d) λ (e) μ (f) σ

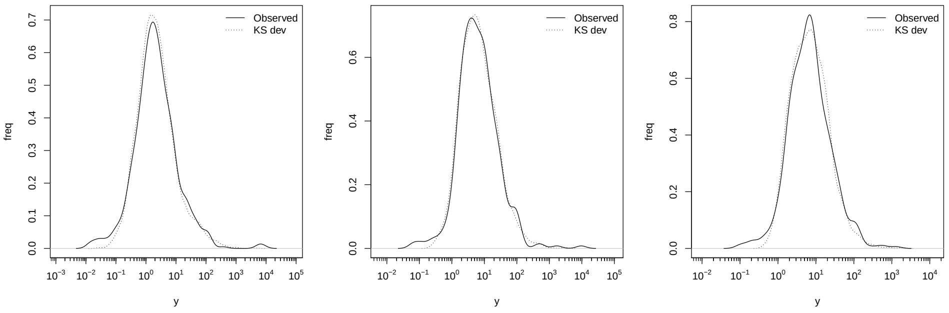
Fig. B1: Prior (gray) and posterior (black) densities for simulated cross–sectional data. Estimated seroconversion rate λ, and the two noise parameters (μ, σ). (a), (b), (c): λ_{sim} = 0.03 (1/yr); (d), (e), (f): λ_{sim} = 0.013 (1/yr). The prior for $\log(\lambda)$ was $N(\log(0.1), 5.0)$; for μ this was $N(0.1, 0.5)$ and for $\log(\sigma)$ $N(\log(0.6), 0.5)$.



(a) λ (b) μ (c) σ

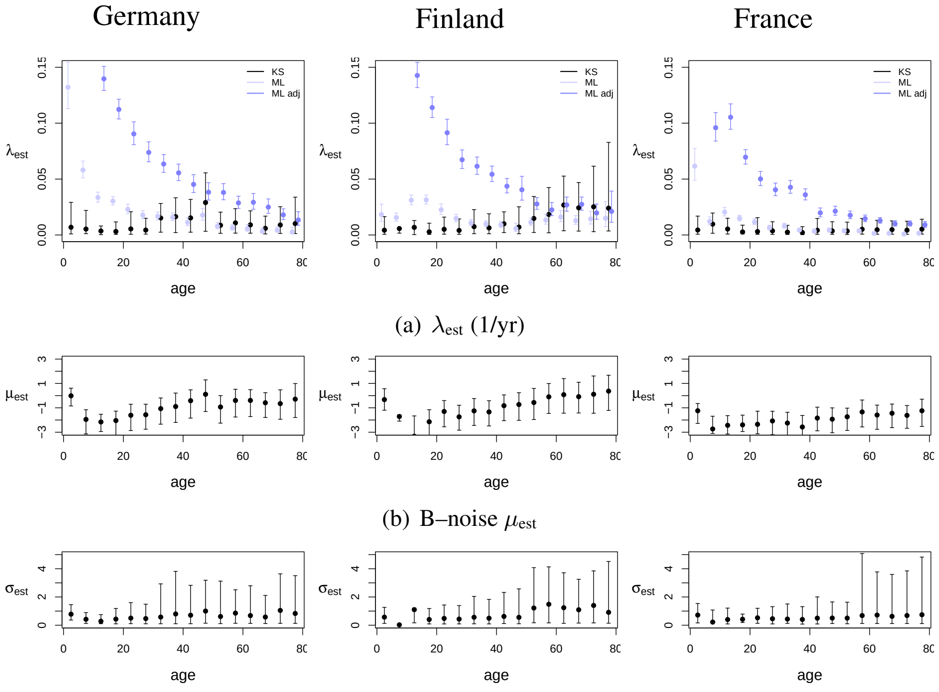
(d) λ (e) μ (f) σ

Fig. B2: Prior (gray) and posterior (black) densities for cross–sectional data from the Netherlands (ESEN study). Estimated seroconversion rate λ, and the two noise parameters (μ, σ). (a), (b), (c): ages 5 – 10 yr; (d), (e), (f): ages 55 – 60 yr. The prior for $\log(\lambda)$ was $N(\log(1.0), 2.0)$; for μ this was $N(1.0, 1.0)$ and for $\log(\sigma)$ $N(\log(1.0), 1.0)$.



(a) Netherlands, 5–10 yr (b) Netherlands, 20–25 yr (c) Netherlands, 55–60 yr

Fig. B3: Densities for population samples of data of the Netherlands from the ESEN study, and densities from samples fitted by matching EDFs.

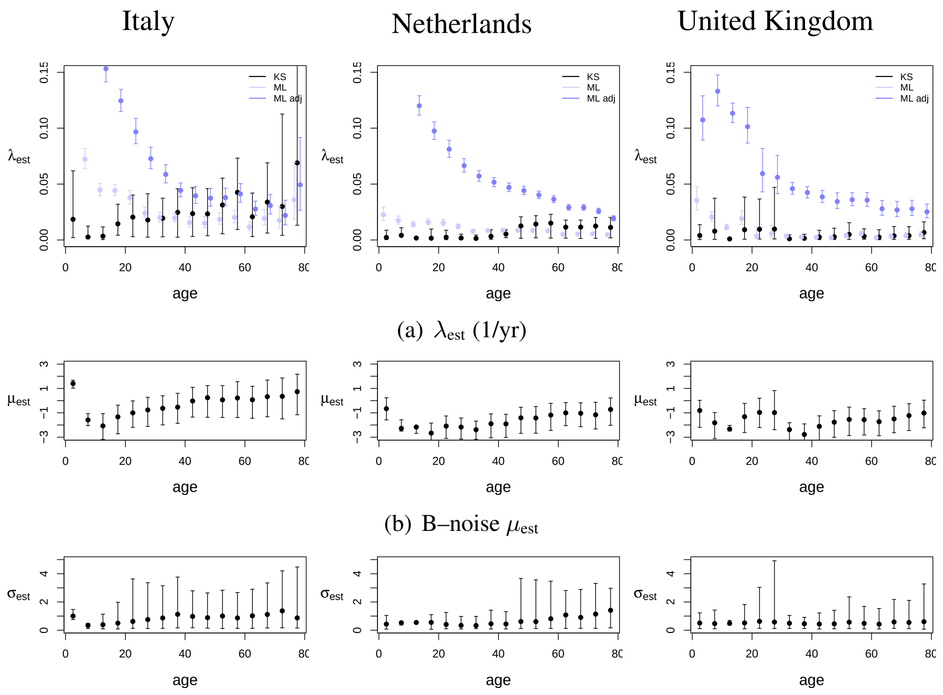


(a) λ_{est} (1/yr)

(b) B–noise μ_{est}

(c) B–noise σ_{est}

Fig. B4: ESEN data, (a) Estimated seroconversion rates λ_{est} by (5 yr) age categories. Two fitting methods are shown. ML: maximum likelihood seroincidence[14] with B–noise fixed $\nu = 3.0$ IU/ml. ML adj: maximum likelihood seroincidence with B–noise adjusted to the 95th percentile of the distribution estimated by EDF. KS: EDF method simulating vaccination at age 2, using p_{KS}, fitted by ABC jointly estimating λ and the two noise parameters. (b) Estimated log mean μ of B–noise. (c) Estimated log sd σ of B–noise.

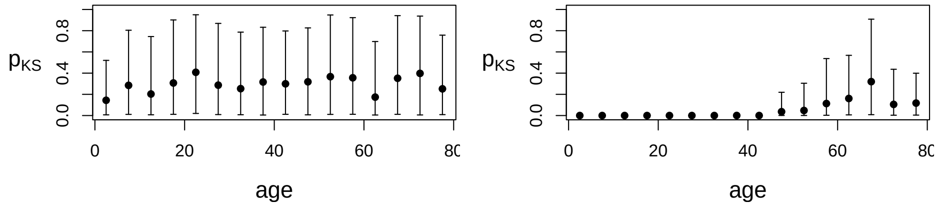


(a) λ_{est} (1/yr)

(b) B–noise μ_{est}

(c) B–noise σ_{est}

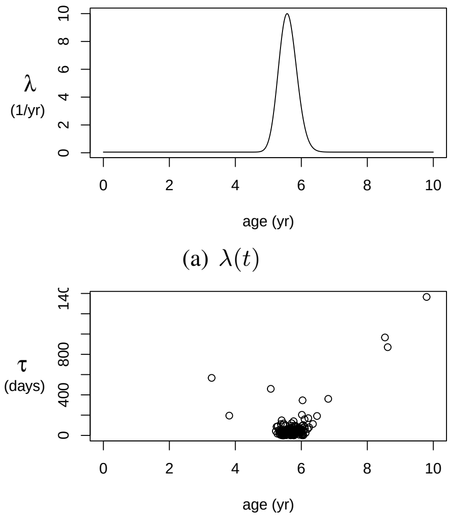
Fig. B5: ESEN data, (a) Estimated seroconversion rates λ_{est} by (5 yr) age categories. Two fitting methods are shown. ML: maximum likelihood seroincidence[14] with B–noise fixed $\nu = 3.0$ IU/ml. ML adj: maximum likelihood seroincidence with B–noise adjusted to the 95th percentile of the distribution estimated by EDF. KS: EDF method simulating vaccination at age 2, using p_{KS}, fitted by ABC jointly estimating λ and the two noise parameters. (b) Estimated log mean μ of B–noise. (c) Estimated log sd σ of B–noise.



(a) No vaccination

(b) Vaccination at age 2

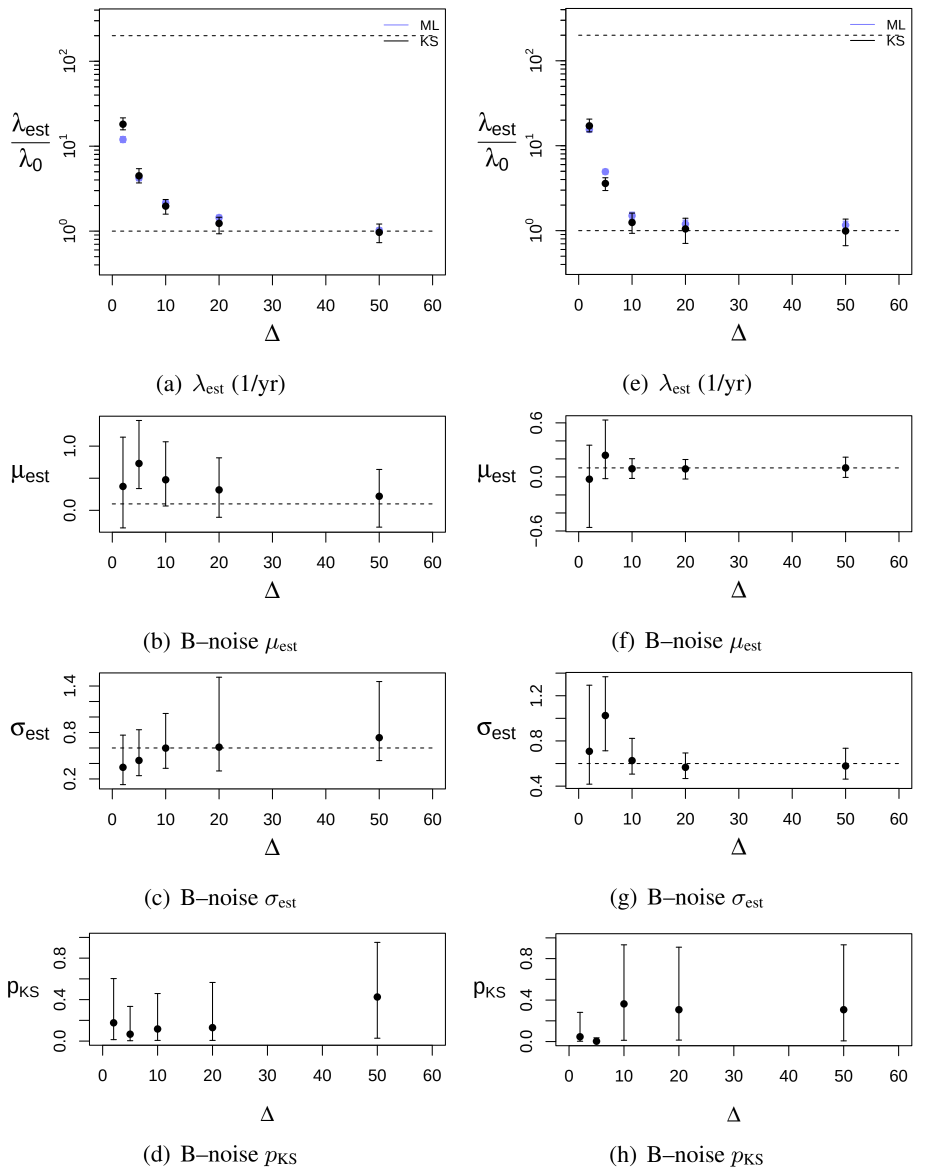
Fig. B6: Kolmogorov probabilities for EDFs of antibody samples generated from posterior parameter estimates in approximate Bayesian computation. Simulated cross–sectional data without (a) and with (b) vaccination at age 2. Observed cross–sectional population data for the Netherlands from the ESEN study for pertussis, analysed in 10 yr age categories, results in Figures 9 and B5.



(a) $\lambda(t)$

(b) τ sampled

Fig. B7: (a) Variation in λ. A subject sampled at age 10 was exposed to an outbreak 5 years ago, causing λ to increase from a baseline 0.05 (1/yr) to a peak infection rate of 10 (1/yr). Duration of the outbreak was 465 days, approximately. (b) sample of intervals for this nonhomogeneous Poisson process (n=25).



(a) λ_{est} (1/yr)

(e) λ_{est} (1/yr)

(b) B–noise μ_{est}

(f) B–noise μ_{est}

(c) B–noise σ_{est}

(g) B–noise σ_{est}

(d) B–noise p_{KS}

(h) B–noise p_{KS}

Fig. B8: Estimated λ_{est} and B–noise parameters ($\mu_{\text{est}}, \sigma_{\text{est}}$) for an outbreak 2, 5, 10, 20 and 50 years ago (Δ) at time of sampling, for subjects aged 0 – 80 years (a–d) or 0 – 10 years (e–h) (uniform age distributions). Baseline infection rate $\lambda_0 = 0.05$ (1/yr), during the outbreak a peak rate of $\lambda_1 = 10$ (1/yr) is reached. Simulated B–noise distribution parameters: $(\mu_{\text{sim}}, \sigma_{\text{sim}}) = (0.1, 0.6)$. ML: maximum likelihood estimation; KS: EDF based method. Note how recent changes in λ cause decreased p_{KS} (d, h).