 Vol 30: 677–698, 2003

Joint Modelling of Recurrent Infections and Antibody Response by Bayesian Data Augmentation

MERVI EEROLA
National Public Health Institute and University of Helsinki

DARIO GASBARRA
University of Helsinki

P. HELENA MÄKELÄ
National Public Health Institute

HENRI LINDEN
University of Helsinki

ANDREI ANDREEV
Swedish School of Economy

ABSTRACT. A joint dynamic model for the interdependence between infection, immunity and risk of disease is presented. Recurrent latent infections are modelled as realizations from a renewal process and antibody dynamics as a diffusion with a decreasing drift modified by the stimulating effect of the random infections. The augmented submodels are estimated simultaneously in one large Markov chain Monte Carlo algorithm. As an example, we consider the risk of recurrent ear infections when having only partially observed information on bacterial carriage and antibody concentrations.

Key words: data augmentation, interdependent random processes, MCMC for conditioned diffusions, pneumococcal ear infections

1. Introduction

Complex biological systems are almost always only partially observable and standard statistical modelling, based solely on observed data, can sometimes result in misleading conclusions about the relationships of interest. Data augmentation, which was proposed initially by Tanner & Wong (1987), has in its various forms proved to be useful when reconstructing the complete system by adding the unobservable part to the likelihood. The joint distribution and its conditionals then often become analytically intractable and numerical methods are needed in estimation. In Bayesian analysis, Markov chain Monte Carlo (MCMC) methods are powerful techniques when building flexible models with latent structures.

In this work, we augment the likelihood to reconstruct a complex interdependent system of random processes. Our motivation comes from infectious diseases where an individual infectious process can never be completely observed. In particular, the time of infection is hardly ever observable. Therefore, the time course from infection to actual disease is often poorly understood and involves complicated interdependencies between microbial virulence factors and immunological processes. Here we attempt to model the risk of a recurrent infectious disease when having only partially observed information on bacterial carriage episodes and antibody level.

Although we concentrate on the analysis of a particular data set, the modelling approach is more general. In many applications, the development of a continuously varying, incompletely observed process is either affected by random events or the process itself affects the risk of some event of interest in time. We shall return to the general methodological connections in section 10.

Preliminary analysis of the empirical problem by standard hazard models suggested complex interactions between the infection pattern and the development of natural antibodies. This is the reason for explicit modelling of the three component processes instead of using the two only as time-varying covariates in the disease model. Predictive calculations, which are often of interest in medical applications, also require probabilistic specification of all parts of the model. Furthermore, we hope to be able to demonstrate that joint modelling of the underlying interacting processes, instead of the observations, allows one to easily integrate data arising from different sources, as in this study.

The structure of the paper is the following: the data and preliminary analysis are described in section 2, and section 3 presents the underlying biological assumptions. In section 4, the structure of the component models and their interdependence is presented and in sections 5 and 6, the priors and the principles of estimation, as well as the posterior distributions are given. Section 7 describes briefly the implementation of the programme and section 8 displays the empirical results. In section 9, sensitivity of the results on modelling and prior assumptions, as well as the model fit is considered. The paper is concluded by discussion in section 10. Details of the MCMC algorithm are given in the appendix.

2. Application and data

Ear infection (acute otitis media, AOM) is one of the most frequent diagnosis of childhood diseases and is most prevalent in children under the age of 2 years. Pneumococcus (Pnc) is one of the major pathogens causing ear infections. Protection induced by current pneumococcal vaccines is limited to the few Pnc serotypes used in the vaccine. Pneumococcal surface protein antigens, which are independent of serotypes, are promising candidates for new vaccines because even young children, who are the primary target group of vaccination, produce antibodies to them. Thus far protectivity of these antibodies has been studied mostly in animals. Understanding of the dynamical relationship between Pnc infections and natural antibodies to Pnc proteins would be helpful in vaccine development. We present a model to analyse the natural development of antibodies in response to latent infections and their potential effect on the risk of the disease, ear infection. For modelling purposes, we consider only the production of antibodies to one pneumococcal protein, pneumococcal surface adhesin A (PsaA), as part of an infectious process.

In the FinOM Cohort Study, 329 children in the Tampere area, Finland, were followed from the age of 2 months until 2 years. The aim was to study the natural course of ear infections and collect information on the risk factors of AOM and the immunological development of the children. During 10 scheduled visits at 2, 3, 4, 5, 6, 9, 12, 15, 18 and 24 months of age the prevalence of asymptomatic Pnc carriage was measured. Blood samples, from which the antibody concentrations were determined, were obtained at 6, 12, 18 and 24 months of age. As a comparison, a blood sample from the mother was obtained at the first visit. The parents were advised to take the child to a study clinic in case of symptoms of (viral) respiratory infection. During these sick visits, carriage of Pnc was measured, and if ear infection was diagnosed, a middle ear fluid sample was collected for aetiological diagnosis of it. For more details of the study design and results, see Kilpi *et al.* (2001) and Vesa *et al.* (2001).

In summary, the data consist of repeated measurements of carriage of Pnc and antibody concentrations at fixed, predetermined time points when the child was healthy, as well as of measurements of carriage and antibodies at random sick visits when either only viral respiratory infection, ear infection, or both, were detected. In our analysis, we use all data, whether from fixed or random time points, as observations from underlying continuous processes. In some cases this is expected to cause observation bias (e.g. for carriage prevalence) which has been accounted for in the modelling. From the modelling point of view, we consider the interdependence of three event processes (latent Pnc infection, viral infection and ear infection caused by Pnc) and the continuously varying process of Pnc protein antibodies.

Figure 1 shows that although the prevalence of Pnc carriage continues to increase up to 24 months of age, the risk of PncAOM does not increase after 12 months. This suggests that the individual susceptibility changes during the first 2 years of life. We shall study this more carefully with the aid of the model. Earlier analysis of these data (Rapola *et al.*, 2000) showed that children with no recorded history of Pnc had low antibody concentrations near to the lowest measurable value in the laboratory tests (detection limit). Whether or not these antibodies reduce the risk of Pnc ear infection and perhaps even pneumococcal carriage is not known and is more difficult to establish. Figure 2 shows all measurements of PsaA antibody concentrations and fitted mean curves to the points for children at the diagnosis of PncAOM and an overall mean curve. Interestingly, children who experienced PncAOM early (before 12 months) had antibody concentrations at the time of diagnosis above the mean level for that age whereas at later occurrences the antibody levels did not differ from the average level.

Terminology. We follow the convention of infectious disease modelling and use the word 'infection' for the onset of asymptomatic carriage of Pnc. For infections with viruses, we

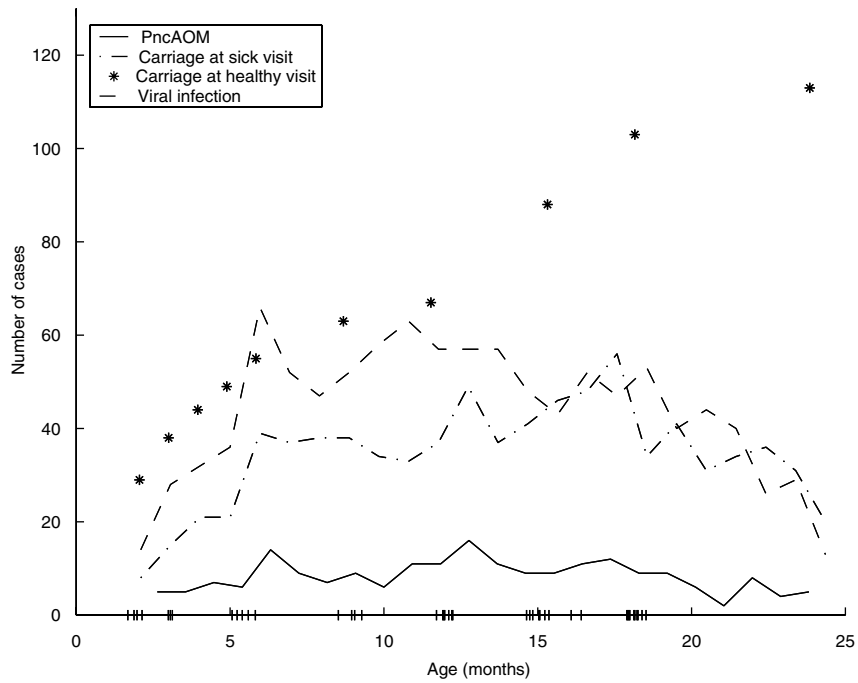

Fig. 1. Observed number of cases of pneumococcal carriage at healthy and at sick visits, viral infection and Pnc ear infection by age (months) (drop-out times are indicated by (+) at the *x*-axis).

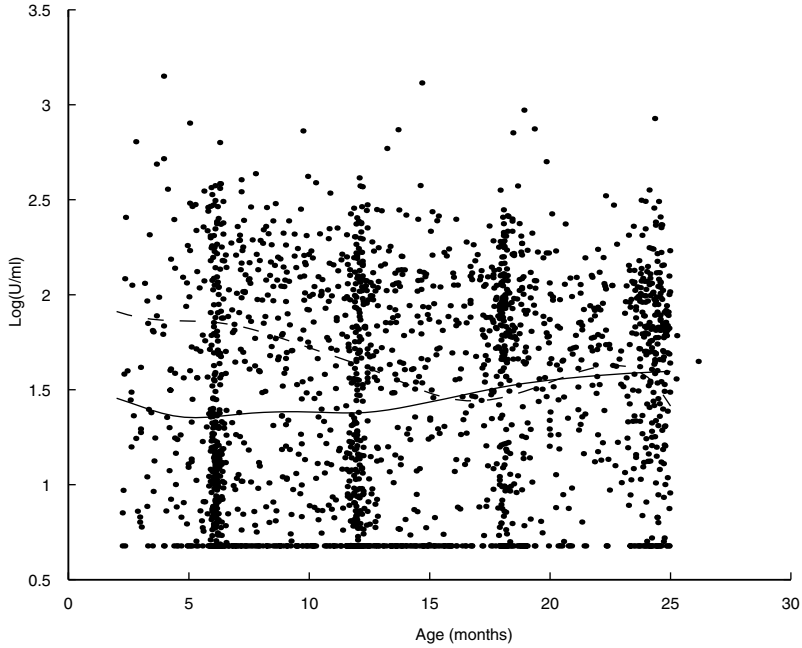

Fig. 2. Anti-PsaA concentrations ($\log_{10}$) at health (6, 12, 18, 24 months) and sick visits. Solid line: Average smoothed level. Dashed line: Concentrations at the diagnosis of PncAOM. Detection limit: 4.76 U/ml.

explicitly write 'viral infection'. We use the word 'disease' or 'ear infection' or 'PncAOM' for the observable clinical infectious disease, here caused by Pnc.

3. Biological assumptions

The modelling is based on the clinical findings that bacterial carriage together with a concomitant viral infection provide the preconditions for AOM. Infection pressure from outside (e.g. siblings, daycare) increases the spread of infection (carriage) among young children but only some of those exposed (carriers) acquire the actual disease, ear infection. Our model assumes that individual proneness to AOM and individual ability to produce protective antibody, which help in eliminating the bacteria, are important determinants which modify the selection probability of the disease in the carriers (Fig. 3).

Dynamic modelling of the underlying infections and antibody level is needed because infections themselves affect the development of antibody production in young children. After losing maternally derived antibodies in a few months after birth, antibody production occurs in response to infections, that is, when encountering the antigen. This process is modified through maturation by age. Each infection therefore acts as an immunogenic stimulus raising the antibody level which is followed by a slow decay in the absence of any new stimulus. The schematic dynamics of the joint model is presented in Fig. 4.

The decay is a composite of several factors but for simplicity we consider them jointly as a common decaying drift of the antibody level. In subsequent infections the immunological memory, especially the memory cells brought about during the first response, will enhance the antibody production. The effect of the stimulus depends on the level of pre-existing antibodies: the higher the level, the smaller the increase. Hence, the infectious process is characterized by

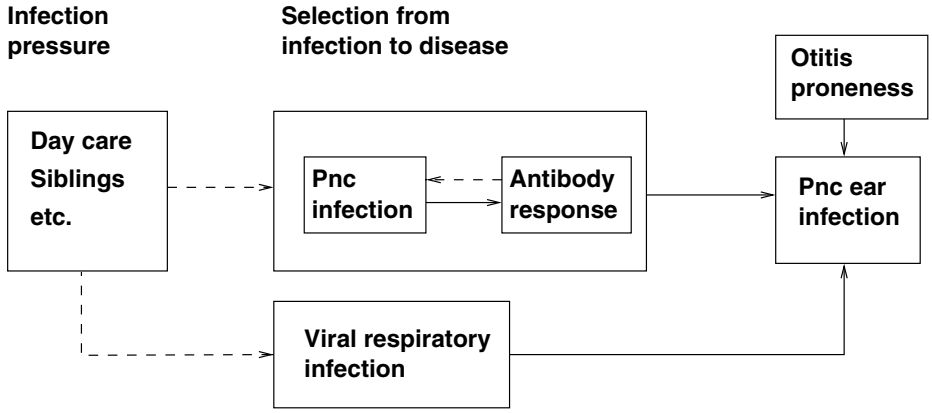

Fig. 3. Interdependency between the model components. Connections marked with dashed line arrows are not implemented in the model

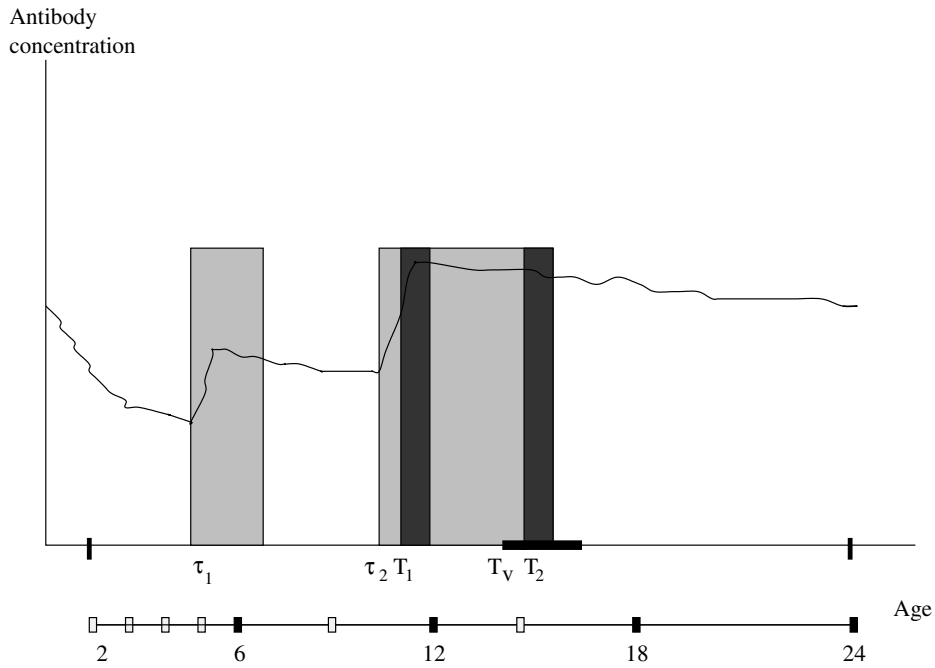

Fig. 4. Schematic presentation of the dynamics of the joint model: recurrent infections at τ_k induce antibody production (solid curve) and positive hazard of AOM during the infection episodes (shaded boxes). Concomitant viral infection (bold line) at T_v increases the hazard of AOM episodes which start at T_1 and T_2 (dark shaded boxes) with incubation times $T_1 - \tau_2$ and $T_2 - \tau_2$. The follow-up is from 2 to 24 months with scheduled observations at 2, 3, 4, 5, 6, 9, 12, 15, 18 and 24 months: carriage (hollow boxes), carriage and serum sample (black boxes).

random infection followed by a humoral immune response, which, if effective enough, should change the risk of the disease in time. It is, however, expected that this dependence is complicated because some children are particularly prone to ear infections due to reasons not

directly related to immunology. The history of previous ear infections is a measure for such 'otitis-proneness'. As the protection induced by antibodies is only temporary, if in this case at all effective, the pattern of carriage – immune response – disease may repeat randomly for an invididual, depending on the development of protective immunity.

The challenge in modelling arises from the fact that the antibody responses to natural infections (indicated in Fig. 4) as well as the natural infections themselves are unobserved. This sets a limit especially to which parts of the antibody process can actually be estimated. The main interest here is in the interaction of the component processes, not in a detailed analysis of the antibodies themselves. We shall return to this in section 4.2 when discussing parameter estimation.

4. Joint model

The joint model for the components of the infectious process is estimated simultaneously by one large MCMC algorithm. Our aim is to combine the different sources of data as effectively as possible. In Bayesian analysis, the submodels are prior information of the system dynamics and observed data are used to restrict the analysis to those realizations of the processes which are supported by the data.

4.1. Model for latent carriage episodes

As the number of natural infections is unknown and must be estimated from the prevalence data, we need to propose models with varying dimensions, that is, with varying number of carriage episodes. The infection episodes, obtained by proposing infection times and durations in the MCMC algorithm, are the basic units in our analysis as we assume that the risk of Pnc ear infection is positive only during simultaneous carriage of Pnc. This relationship has been shown earlier (e.g. Harper, 1999) and also in our data.

We model the 10 measurements of Pnc carriage as prevalence observations of an underlying binary renewal process (Y_t) taking values 0 (non-carrier) or 1 (carrier) with jumps of $+1$ or -1 at the random times of acquiring and clearing carriage. Neither the times of infection nor the duration of carriage are observed and the likelihood of the carriage prevalence data is augmented by these random variables. Typically, we observe a series like '0, 0, 1, 0, 1, 1, 0, 0, 0, 1' and their measurement times for each individual, allowing for intermediate missing values. As a child must have been a carrier of Pnc when PncAOM or carriage at a sick visit was detected, we use this information as additional points of prevalence in the estimation. However, as it is suspected that Pnc carriage is more prevalent at sick visits than at health visits, direct inclusion of sick visit observations might cause selection bias in carriage estimation. We correct for this by modelling the intensity of a sick visit as a sum of two intensities

$$\lambda^S(t) = \lambda^+ Y_t + \lambda^-(1 - Y_t) \tag{1}$$

where λ^+ is the intensity of sick visit when the child is a Pnc carrier and λ^- is the intensity of sick visit when not. We denote the process counting sick visits by $S = (S_t)$.

The latent process of carriage episodes is a similar mixture of intensities. Let $\{\tau_k\}$ be the sequence of latent infection times. For given t, define $\tau(t) = \max\{\tau_k : \tau_k \le t\}$ as the last infection time before t. The stochastic characteristic of the process (Y_t), which depends on the changes at the infection and clearing times, as well as on the changes in the status of the process, is then

$$A(\mathrm{d}x, \mathrm{d}t) = (1 - Y_{t-})\delta_{+1}(\mathrm{d}x)\lambda_{01}(t)\mathrm{d}t + Y_{t-}\delta_{-1}(\mathrm{d}x)\lambda_{10}(t - \tau(t))\mathrm{d}t, \tag{2}$$

where $\delta_{\cdot}(\mathrm{d}x)$ is the jump of size ± 1 at the times of acquiring and clearing carriage. $\lambda_{01}(t)$ is the age-dependent intensity of a new infection, given that the child is not a carrier. It is a piecewise constant function which is estimated non-parametrically. The clearing intensity $\lambda_{10}(t - \tau(t))$ controls the random duration of carriage d. As there is no *a priori* connection between clearing times and antibody level, we define the model in terms of the pairs of infection times and durations $\{(\tau_k, d_k)\}$. When augmenting the prevalence data by them, the likelihood becomes simply an indicator which needs to match with the values of the underlying process Y at the observation times.

According to the study protocol, the children were taken to the study clinics in case of symptoms of respiratory infection. Thus, if viruses were detected at the sick visit, the onset times of viral respiratory infections were considered as observed data, unlike the onset times of asymptomatic Pnc infection. It has been found in clinical studies that AOM occurs soon after clinical symptoms of viral respiratory infections (Ruuskanen & Heikkinen, 1994). Therefore, if viral infection occurs within an estimated Pnc carriage, it is taken as a risk factor of PncAOM, otherwise not. We denote viral infections by $V = (V_t)$.

4.2. *Model for antibody dynamics*

The decay in antibody level is interrupted by random infections which stimulate the production of new antibody. Thus, we should specify a continuous process whose development is conditioned by random events. Brownian motion with a drift is a simple continuous process which is characterized by the drift governing its mean behaviour and random variation around the mean which is the diffusion coefficient. Hence, we assume that relative changes in the antibody level can be represented by the following stochastic differential equation

$$\mathrm{d}X_t = X_t(\sigma\,\mathrm{d}W_t + \rho_t^*\,\mathrm{d}t), \tag{3}$$

where W_t is a Wiener process and σ a diffusion coefficient around the drift ρ_t^*. This random variation can be interpreted as interindividual variation. At the infection times τ_k, the common decaying drift ($\rho < 0$) is modified to account for the stimulus in antibody production

$$\rho_t^* = \rho + \sum_k 1_{\{\tau_k < t \le \tau_k + c\}} (X_{t-})^{-1} \psi_k. \tag{4}$$

According to previous human immunogenicity studies, the time to maximum response takes about 2 weeks so the drift increases within an interval of $c = 14$ days. The increase at τ_k is assumed to be inversely related to the current level of antibodies which reflects the understanding that some saturation level exists. The jump has a random coefficient ψ_k which allows for intra- and interindividual variation in antibody response. As the distribution of antibody concentrations is highly skewed to the left, we work with log-antibody concentrations. When using the same notation X for log-antibodies, the model reduces to the form

$$\mathrm{d}X_t = \sigma\,\mathrm{d}W_t + \rho_t^*\,\mathrm{d}t. \tag{5}$$

Otherwise the antibody dynamics is based on the following assumptions, several of which are simplifications of the complex biological events behind them:

- At birth the child receives approximately 60% of the mother's IgG antibody concentration which is used as the initial value (X_0) of the antibody process.
- The common drift, which is not related to the jumps, and the diffusion coefficient, are estimated *a priori* from the antibody data of children who had no observed Pnc contacts ($n = 62$), and thus were likely to have no jumps either. The average decay of antibodies is -0.005 log-units per day (median $= -0.002$), and this is used as an estimate of the drift ρ.

The estimate of the standard deviation around the drift is $\hat{\sigma} = 0.02$. These fixed values of ρ and σ are used in the MCMC algorithm.

- The enhancing effect of the immunological memory due to earlier infections is only implicitly measured by the pre-existing level of antibodies. Explicit modelling would require letting the response to depend on the history of the infection process Y, i.e. on the number of previous infections, but it is not done here.
- Given the data, the posterior antibody process is constrained to match with the observed log-antibody concentrations. If the observed concentrations are censored below (i.e. have value at the detection limit 4.76 U/ml), the posterior process is constrained to have values below the detection limit.

4.3. Model for ear infections

The model for ear infections should capture the effect of selection from exposed (infected with Pnc) to diseased (PncAOM) which was characterized in section 3. We, therefore, define a multiplicative intensity model for ear infections as a function of a common age-dependent baseline intensity $\mu_0(t)$, current level of PsaA antibodies, concomitant viral infection, and history of previous Pnc ear infections. The intensity of ear infection is linked to the infection process Y by being positive only when the child is infected with Pnc (i.e. when $Y_t = 1$). On the basis of Fig. 2, we allow the regression coefficient α_t of PsaA antibodies to be age-dependent. Both $\mu_0(t)$ and α_t are piecewise constant functions and were estimated in the same way as the infection intensity $\lambda_{01}(t)$ (appendix (A3)). The effect of an overlapping viral infection episode (i.e. when $V(t) = Y(t) = 1$), is estimated by β_1 and the effect of the number of previous PncAOM N_{t-} by β_2. The intensity model of PncAOM is then of the form

$$\mu(t) = Y_t \mu_0(t) \exp((\alpha_t X_t / x^*) + \beta_1 V_t + \beta_2 N_{t-}), \tag{6}$$

where $x^* = 1.56$, the logarithm of the detection limit indicating no observable antibodies. According to the study protocol, the children were given antibiotic treatment if AOM was detected. As this affects the duration of the AOM episode (but, according to Gray *et al.*, 1980, not the duration of carriage), its duration was not estimated in the model. Moreover, in order not to model control visits as new AOM events, we used the rule that a new (PncAOM) episode can start only if at least 30 days have elapsed since the previous diagnosis of PncAOM. This convention has been used in many studies on ear infection (e.g. Kilpi *et al.*, 2001).

5. Prior specifications

It is clear that all assumptions in the submodels, either functional or statistical, are specifications on some prior knowledge. All results therefore depend, more or less, on their validity. We shall return to this in section 6.

Piecewise constant parameter processes. α_t, $\mu_0(t)$ and $\lambda_{01}(t)$ are modelled non-parametrically as piecewise constant processes (cf. Arjas & Gasbarra, 1994) with the structure

$$f(t) = \sum_{i=1}^{N_T+1} f_i I_{[R_i, R_{i-1})}(t)$$

where $R_0 = 0$, $R_{N_T+1} = T$. The change points $\{R_1 \leq R_2 \leq \cdots R_{N_T}\}$ form independent Poisson processes over the interval $[0, T]$, where T is the maximum observation time of all individuals. We need to define the distribution of the initial level f_1, the intensity measure $\eta(\mathrm{d}t)$ of the Poisson process of the change points, and the transition kernel for the successive levels f_i,

$i > 1$. For the successive levels, we use a Gaussian–Markov transition kernel with variance proportional to the length of the interarrival times: $[f_{i+1}|f_i, R_i, R_{i-1}] \simeq \mathcal{N}(f_i, \sigma^2(R_i - R_{i-1}))$. The cumulative intensity $\eta([0, T])$ gives the expected number of the change points. This together with the variance in the transition distribution is used to tune *a priori* the 'smoothness' of the random function $f(t)$. A non-uniform intensity for the change points is used to get a finer resolution for the beginning of the observation interval where the follow-up is more intensive. This correspond to a non-uniform choice of the bandwidth in non-parametric kernel estimation. All of the piecewise constant functions α_t, $\mu_0(t)$ and $\lambda_{01}(t)$ have the same prior values, given below, except for the parameters of the initial level distribution.

Carriage. Previous studies on Pnc carriage indicate that the duration of carriage is highly variable (1–17 months) and serotype-specific (Gray *et al*., 1980; Smith *et al*., 1993). Previous analyses with some parts of the same data set as ours, concerning only carriage data, estimated the average duration of carriage to be 19 days (Ekholm *et al*., 2001, using maximum liklihood-estimation) and 70 days for the most prevalent serogroups 6, 19 and 23 (Auranen *et al*., 2000, using data augmentation in MCMC simulation). Here we specify the duration d be a Gamma(1, v)-variate, where v is a hyperparameter having a distribution Gamma(2, 90). This yields a prior mean of 64 days and median 30 days.

When estimating the infection intensity $\lambda_{01}(t)$, the intensity of the change points is let to be of the form $\eta(\mathrm{d}t) = \frac{5}{100}\exp(-1/500t)\mathrm{d}t$. This prior corresponds to five expected change points in the interval $[0, T]$. The distribution of the initial level is $\log \lambda_{01}(1) \simeq \mathcal{N}(-5,9)$, which corresponds to $\lambda_{01}(1) \approx 0.007$, a rough starting value estimated from Fig. 1, where we have on average 50–60 infected individuals from the total 329 in 22 months. The distribution of the successive loglevels is $[\log \lambda_{01}(i+1)|\ \log \lambda_{01}(i),\ S_i,\ S_{i-1}] \simeq \mathcal{N}(\ \log \lambda_{01}(i),\ \sigma^2(S_i - S_{i-1}))$, with $\sigma = 0.01$.

Sick visits and viral infections. As these events are observable and quite common (Table 1), the prior mean rate will not matter very much. We specify a prior model of exp(1/100) for all of them and let the data decide whether or not $\lambda^+ > \lambda^-$. This prior model yields approximately 3.5 events per year.

Antibody level. Having estimated ρ and σ *a priori*, the only parameter needed to sample is the jump size ψ_k at the infection times. The prior of Gamma(2, 0.5) favours small jumps but has a long tail giving some probability even for large jumps.

Table 1. *Number of cases and samples in the data*

Type of finding	Cases	Total	%
Pnc carriage			
Carrier at healthy visit at least once	238	329	72.3
Positive samples at healthy visits	649	3026	21.4
Carrier at sick visit at least once	223	329	67.8
Positive samples at sick visits	825	2007	41.1
Viral findings			
Children with at least one	254	329	77.2
Positive samples at sick visits	837	2007	41.7
Number of episodes (30 days apart)	761		
PncAOM			
Children with at least one	109	329	33.3
Children with only one	65	109	59.6
Number of events	201		
Number of episodes (30 days apart)	175		
Children with incomplete follow-up	49	329	14.9

Ear infection. The prior for the regression parameters β_1 and β_2 is $\mathcal{N}(0, 9)$. For the time-dependent regression parameter α_t, we set $\alpha_1 \simeq \mathcal{N}(0, 9)$ and for the common baseline intensity, the initial log-level is $\log \mu_0(1) \simeq \mathcal{N}(-5, 9)$. The intensity of the change points and transition distribution are the same as for $\lambda_{01}(t)$.

6. Joint density and full conditional distributions

The joint density comprises models for the observations, the underlying augmented processes and the parameters. For brevity, we will use the following notation: for sequences of random variables or random processes we simply write, for example, $\tau = \{\tau_k\}$ for infection times, $Y^{(i)} = \{Y_t^{(i)},\ t \in [0, T_i]\}$ for infection process, and $y^{(i)} = \{y_{s_j}^{(i)},\ j = 1, \ldots, n_i\}, x^{(i)} = \{x_{t_k}^{(i)}, k = 1, \ldots, m_i\}, N^{(i)} = \{\Delta N_t^{(i)}, t \in [0, T_i]\}, V^{(i)} = \{\Delta V_t^{(i)}, t \in [0, T_i]\}$ and $S^{(i)} = \{\Delta S_t^{(i)},\ t \in [0, T_i]\}$ for carriage, antibody level, AOM, viral infection and sick visit observations, respectively. Given the model parameters $\boldsymbol{\theta}$, the individual processes are independent. In the following, we therefore drop the superscript i for individuals. The joint density of the parameters, unobservables and data can then be factorized as

$$p(Y, S, X, N, V, y, x, \boldsymbol{\theta}) = p(y|Y)p(x|X)p(Y|\boldsymbol{\theta})p(S|Y, \boldsymbol{\theta})p(X|Y, \boldsymbol{\theta})p(N|Y, X, V, \boldsymbol{\theta})p(V|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

where $p(y|Y) = \prod_j I\{Y_{s_j} = y_{s_j}\}$ and $p(x|X) = \prod_k I\{X_{t_k} = x_{t_k}\}$ are the likelihoods of observed carriage and antibody data given the augmented processes Y and X, and $\boldsymbol{\theta}$ is the vector of model parameters.

The MCMC algorithm, described in the appendix, requires all full conditional distributions of the unknowns. In what follows, we describe in detail the posterior conditional distributions of latent infections and antibody concentrations which involve all other parts of the joint model.

6.1. Posterior of latent infection process

The latent infection process Y is central in our analysis. When the proposed infection intervals change, they affect the dynamics of almost everything in the joint model. Therefore, the posterior density of Y involves all other parts in the model also. It is, conditionally on the prevalence data pattern, sick visits, antibody and AOM processes, proportional to the following product

$$p(Y|\lambda_{01}, d)p(S|Y, \lambda^+, \lambda^-)p(X|Y, \psi)p(N|Y, X, V, \mu_0, \alpha_t, \beta_1, \beta_2) \prod_j I(Y_{s_j} = y_{s_j}). \tag{7}$$

The density of the latent infection process is a product of the Poisson density of becoming infected and the Gamma density of duration

$$\begin{aligned} p(Y|\lambda_{01}, d) = \exp\Bigg(&\log(\lambda_{01}) \int_0^T (1 - Y_{s-})\mathrm{d}Y_s - \lambda_{01} \int_0^T (1 - Y_{s-})\mathrm{d}s\Bigg) \\ &\times \prod_k \frac{v^\alpha}{\Gamma(\alpha)} (d_k - c)^{\alpha-1} \exp(-v(d_k - c)) \end{aligned} \tag{8}$$

where $\int_0^T (1 - Y_{s-})dY_s$ is the number of times the child becomes infected at some $s \in [0, T]$ when being susceptible at $s-$. The density of sick visits is

$$\begin{aligned} p(S|Y, \lambda^+, \lambda^-) = \exp\Bigg(&\log(\lambda^+) \int_0^T Y_{s-}\,\mathrm{d}S_s - \lambda^+ \int_0^T Y_{s-}\,\mathrm{d}s \\ &+ \log(\lambda^-) \int_0^T (1 - Y_{s-})\mathrm{d}S_s - \lambda^- \int_0^T (1 - Y_{s-})\mathrm{d}s\Bigg). \end{aligned} \tag{9}$$

where $\int_0^T Y_{s-}\,\mathrm{d}S_s$ is the number of sick visits in [0, T] when the child is a carrier and $\int_0^T (1 - Y_{s-})\,\mathrm{d}S_s$ when not.

The contribution from the underlying antibody process comes from the interval of increasing drift due to infection at τ_k

$$p(X|Y,\psi) \propto \prod_{t\in(\tau_k,\tau_k+c]} \exp\left(-\frac{(\Delta X_t - (\rho_t^* - \rho_t)\Delta t)^2}{2\sigma^2\Delta t}\right). \tag{10}$$

Finally, the likelihood contribution from the AOM data in the infection interval is

$$p(N|Y,X,V,\mu_0,\alpha_t,\beta_1,\beta_2) = \exp\left(\int_{\tau_k}^{\tau_k+\tau_{k+1}} \log(\mu(s))\mathrm{d}N(s) - \int_{\tau_k}^{\tau_k+\tau_{k+1}} \mu(s)\mathrm{d}s.\right) \tag{11}$$

6.2. *Posterior of antibody level process*

The discretized antibody process evolves as

$$[X_t|X_{t-1}] = X_{t-1} + \rho_t^* + \sigma\epsilon_t \tag{12}$$

with $\epsilon_t \sim N(0, 1)$. This implies that the posterior of X_t is the mean of the densities of the neighbouring values X_{t-1} and X_{t+1} and is proportional to

$$\begin{aligned} p(X_t|X_{t-1},X_{t+1},\psi,Y,N) \propto \exp&\left(-\frac{1}{\sigma^2}\left(X_t - \frac{X_{t-1} + X_{t+1} + \rho^*(t+1) - \rho^*(t)}{2}\right)^2\right) \\ &\times p(N|Y,X,V,\mu_0,\alpha_t,\beta_1,\beta_2)\prod_k I\{X_{t_k} = x_{t_k}\}. \end{aligned} \tag{13}$$

The conditional distribution of the marks (ψ_k) given the corresponding infection times (τ_k) and the antibody process is proportional to the prior gamma density Gamma(2, $\frac{1}{2}$) times the likelihood of X_t.

7. Implementation and convergence

The details of constructing the MCMC algorithm are given in the appendix. As the complexity of the model results in rather long run times, we chose to compare independent parallel runs starting from different random points in the parameter space. The convergence was checked by tests in R-Coda, in particular by the test proposed by Gelman & Rubin (1992) for parallel runs. Three parallel runs of size 50,000 with thinning of 5 and burn-in of 10,000 were sufficient to reach stationarity according to the tests. Most of the parameters stabilised well before 50,000 iterations.

8. Results

We concentrate on results where explicit modelling of the latent infections and antibody level is needed. In most cases, the interest is in the time-dependent (developmental) aspects of the processes, and these results are illustrated in figures. As the storage of all simulation results would have been practically impossible, we shall not show credible intervals for all estimated processes. The posterior means of time-independent population parameters are presented in Table 2. In all cases, the results are of the combined three parallel runs.

Table 2. *Posterior mean, SD, median and 90% credible interval of time-independent population parameters: $\lambda^{+/-}$, intensity (1/1000) of sick visit when Pnc carrier (+) and when not (−); β_1, regression parameter of viral infection (AOM-model); β_2, regression parameter of number of previous PncAOM (AOM model); incubation time, time from infection to disease*

Parameter	Mean	SD	Median	90% CI
λ^+	20.2	7.89	20.2	(18.9, 21.5)
λ^-	6.5	1.91	6.5	(6.2, 6.8)
β_1	1.76	0.22	1.77	(1.37, 2.10)
β_2	−0.01	0.30	−0.02	(−0.50, 0.41)
Duration of carriage (days)	33	0.98	23	(11, 39)
Incubation time (days)	21	1.26	14	(1, 62)

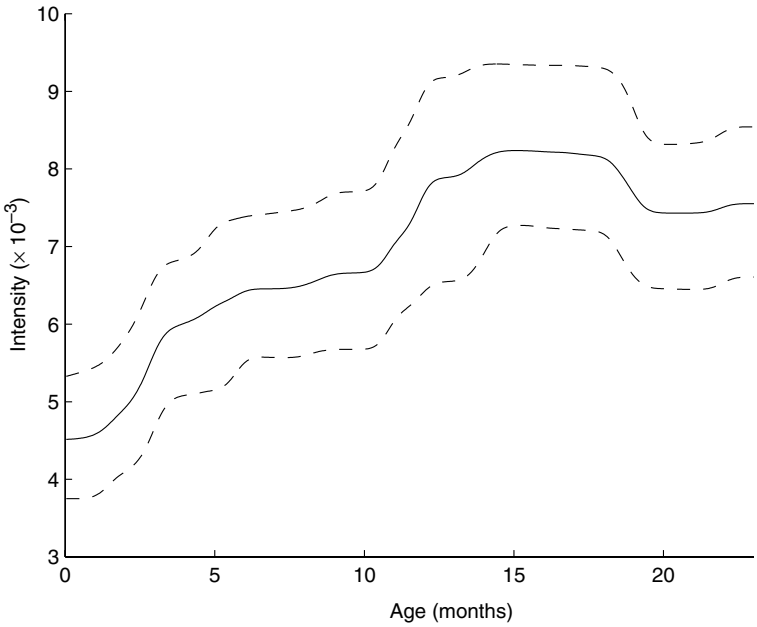

Fig. 5. Posterior intensity ($\lambda_{01}(t)$) of Pnc infection (with pointwise 90% CL).

Infection, duration and incubation time. The posterior intensity of infection, shown in Fig. 5, increases rapidly and levels off at around 18 months of age more clearly than the prevalence data suggests. The mean duration of carriage is 33 days, but Fig. 6 shows that most of the carriage episodes are short. Having estimated the infection episodes, it is possible to estimate both λ^+ and λ^- and conclude that Pnc carriage is likely to be involved with sick visits: the intensity λ^+ is three times higher than λ^-. The correction for the carriage estimation introduced in section 4 is therefore needed. Figure 6 shows also that in most cases PncAOM occurs soon after becoming infected with Pnc but in some cases the incubation time can be rather long, most likely when multiple ear infections occur within the same infection episode.

The mean age of first carriage episode is around 8 months (Table 3). Although the prevalence of carriage doubles after 12 months of age (Fig. 1), children who have several carriages, tend to have them very early: the mean age of sixth carriage episode is 13 months whereas the mean age of second carriage is 15 months.

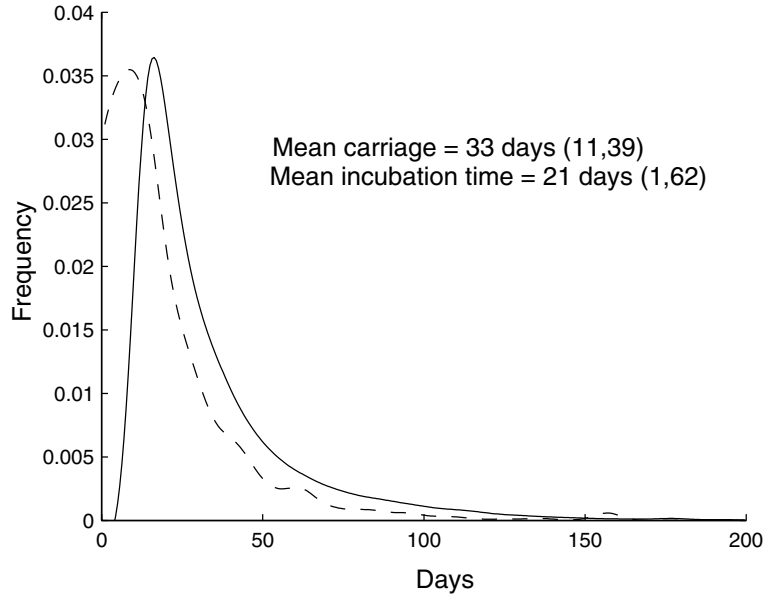

Fig. 6. Posterior distributions of carriage duration d_k (solid line) and incubation time ($T_{\mathrm{AOM}} - \tau_k$) from Pnc infection to PncAOM (dashed line).

Table 3. *Mean age of infection by order of episode, posterior probability of at least n episodes, and estimated number of cases in each order category*

Order of episode	1	2	3	4	5	6
Age (months)	8.4	15.5	15.1	14.5	13.3	13.0
Probability	0.81	0.74	0.60	0.46	0.34	0.25
n	266	245	196	150	112	81

Antibodies to PsaA and infections. The pointwise geometric mean concentrations (GMCs) of antibodies to PsaA in Fig. 7 show that after the decline of inherited antibodies there is a gradual increase, especially from 15 to 18 months of age, which then levels off.

Figure 8a, which displays the duration of carriage episodes by age of infection, indicates that most of the time the duration remains rather close to the mean of 33 days. In the end of the follow-up, from 20 to 24 months, it drops sharply to about half of the mean suggesting a change in the underlying factors. Two factors seem relevant here: a large increase both in carriage prevalence around 15 months (Fig. 1) and in the concentration of anti-PsaA antibodies around 18 months (Fig. 7).

We did not explicitly model the effect of PsaA antibodies on the risk of infection but it is possible to estimate the level of antibodies at the times of infection in the model. Figure 8b shows a high overall level with an increasing trend at 24 months of age which is almost mirror-like compared with the duration of carriage in Fig. 8a. This suggests a possible causal relationship: increasing prevalence of carriage would lead to an increased concentration of anti-PsaA antibodies. This in turn could lead to enhanced elimination of the carried bacteria. Unfortunately there is too little real knowledge of the determinants of carriage duration to consider this more than an intriguing possibility.

Antibodies to PsaA and risk of PncAOM. Figure 2 indicated that the relationship between anti-PsaA antibodies and PncAOM may vary during the follow-up. This was indeed the case

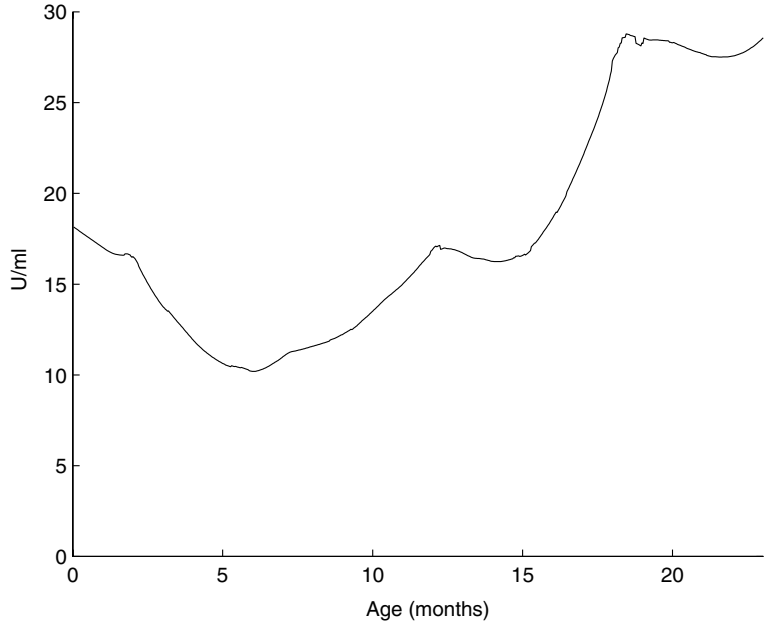

Fig. 7. Natural development of antibodies (X_t) to PsaA. Geometric mean concentration by age.

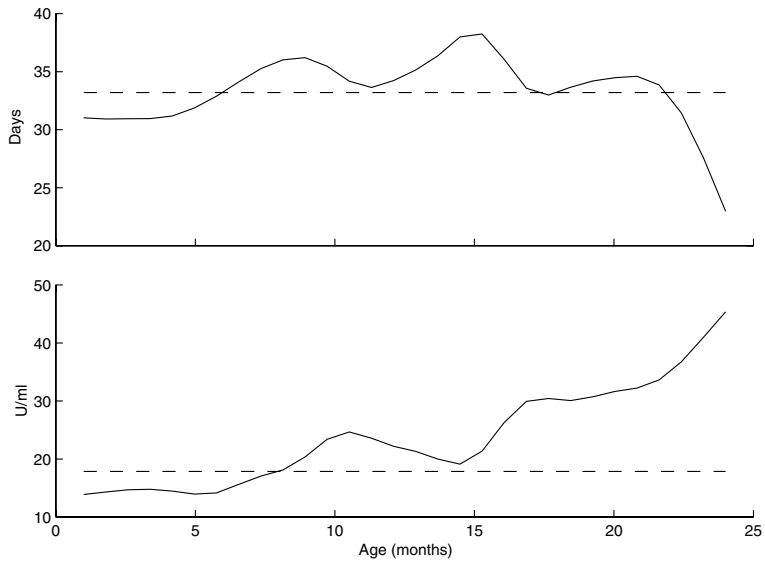

Fig. 8. (a) Duration of carriage (d_k) by age (month) of infection. Dashed line: average duration. (b) Geometric mean of anti-PsaA concentrations at age of infection. Dashed line: overall GMC.

when the regression coefficient α_t of the antibody level (relative to the detection limit) was allowed to depend on age (Fig. 9). During the first year of life, high level of PsaA antibodies seems to increase the risk of PncAOM. It is, however, more probable that increased risk is associated with an environment of high prevalence of Pnc infection and consequent production of PsaA antibodies. During the second year the effect of anti-PsaA turns to negative

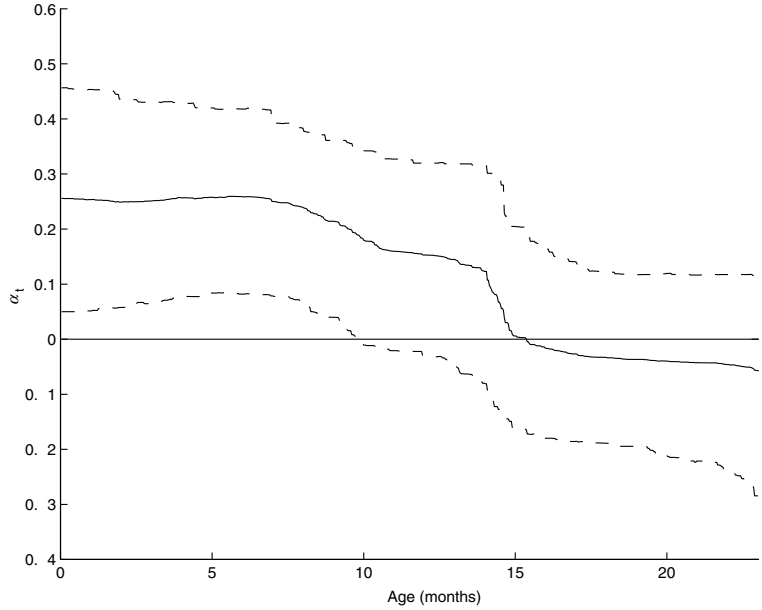

Fig. 9. Age-dependent effect (α_t) of anti-PsaA antibodies on the risk of PncAOM (with pointwise 90 per cent CL).

suggesting a developing protective effect. The posterior credible intervals are, however, rather wide and include the probability of no protective effect at all.

If the age-dependent effect of anti-PsaA in Fig. 9 is indeed real and indicative of a causal relationship, the question of the biological mechanism becomes relevant and intriguing. An immunological explanation, suggesting that antibodies in the older children would be more protective, is not implausible as they would likely be the result of multiple stimulations and immunological memory. This would also be consistent with the increasing antibody concentration occurring at about the same age (Fig. 7). On the other hand, we have only studied the effect of anti-PsaA antibodies whereas the same stimulus by Pnc is expected to lead to the production of antibodies to other Pnc proteins as well. Thus the effect of the anti-PsaA could actually be mediated by some other antibodies whose presence would correlate with anti-PsaA.

Carriage, viral infection and risk of PncAOM. Having estimated the latent infection episodes, we obtain easily posterior probabilities of certain events of interest. In these data, the posterior probability that a healthy carrier acquires clinical symptoms during a carriage episode was as high as 0.5. The probability that viruses are detected during Pnc carriage was 0.27. According to the present knowledge about the pathogenesis of AOM, this means that in every fourth carriage episode there was an elevated risk for the development of otitis media. However, the posterior probability of at least one PncAOM during a carriage episode was only 0.13, indicating that, on average, every seventh carriage resulted in AOM. The large regression coefficient β_1 in Table 2 indicates that viral infection clearly increases the risk of PncAOM but the posterior probability of a concomitant viral infection within episodes where PncAOM occurred, was only 0.58. This is likely due to insufficient sensitivity of the methods used for viral diagnosis. In any case, the synergistic effect of viruses and Pnc is an important part of the pathogenesis of PncAOM.

9. Model assessment

Effect of model assumptions. The basic submodels have a simple structure: a non-parametric intensity for infections, a diffusion model with a common constant drift and diffusion coefficient for antibody level, and a non-parametric intensity for ear infections. The functional assumptions (modified drift at the time of infection, risk of AOM depending on the infection episodes and multiplicative dependence on the antibody level, concomitant viral infection and previous PncAOM history) join the submodels and introduce interindividual heterogeneity which is completely absent in the basic submodels. Although partly unobservable, and thus unverifiable from the data, these assumptions are to our understanding reasonable and almost minimal from the biological point of view. More complex feedback relations between the submodels could in principle be incorporated into the model (e.g. by allowing the antibody level influence also the risk of infection), but the sparsity of the data did not allow for such extensions.

Sensitivity of the submodels was tested by estimating the infection intensity and duration without the other models. Although there is no feedback from the AOM and antibody models to the carriage model, the additional information on Pnc ear infections changed the estimated mean carriage duration from 61 to 33 days. As Pnc ear infections are additional information on the prevalence of Pnc carriage, the joint model is expected to estimate carriage duration more accurately. The effect of PsaA antibodies was studied by standard piecewise hazard models which gave some indication of age-dependency, but in our joint model the whole pattern of the age-dependent coefficient could be obtained.

Effect of prior specifications. Extensive tests with different prior values for the main parameters indicated that the model was insensitive to prior specifications. As the values of the drift ρ and diffusion coefficient σ were fixed (estimated) *a priori*, it was of interest to test whether large perturbations in these values would change the results. A 10-fold value for ρ, compared with its *a priori* estimated value, increased the level of α_t to some extent, but the age-dependent shape remained the same. Other parts of the model were insensitive to substantial changes in ρ and σ. The model was also robust against different prior values of carriage duration d. Although the prior model (different variances in $N(0, \sigma^2)$) for the regression parameters β_1 and β_2 in the ear infection model had no effect on the results, careful tuning of their proposal variances was needed.

Model fit. The latent parts of the joint model, augmented by missing variables, complicate model assessment because obviously it should be based on some observable quantity. Figure 10 shows the overall fit (TTT-plot) of the ear infection model in terms of the AOM counting process N_t and its cumulative posterior intensity $\int \mu_s \, \mathrm{d}s$, the latter of which depends on both of the latent processes X and Y. The joint model slightly underestimates the number of PncAOM before the age of 1 year but after that fits very well.

In Bayesian analysis, model assessment is often based on predictive distributions. The probabilistic agreement of replicated or new data with the estimated model can be studied in this way. In our case, the relevant measures are the following individual predictive probabilities of ear infection

$$P(N_i(\tau_k + d_k) - N_i(\tau_k) \in I | \mathscr{F}^i_{\tau_k}) \tag{14}$$

where $I \subset \{1,2,\ldots\}$ within the infection intervals $(\tau_k, \tau_k + d_k]$. Alternatively, in Fig. 10, the cumulative predictive intensity could be used instead of the estimated one. The conditioning history $\mathscr{F}^i_{\tau_k} = \sigma(H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,\tau_k]})$ is a union of the complete observed history of all other individuals except the ith in the whole follow-up interval $[0, T]$, and the observed history $H^i_{[0,t^*]}$ of the ith individual up to some prediction time t^* (cf. Arjas &

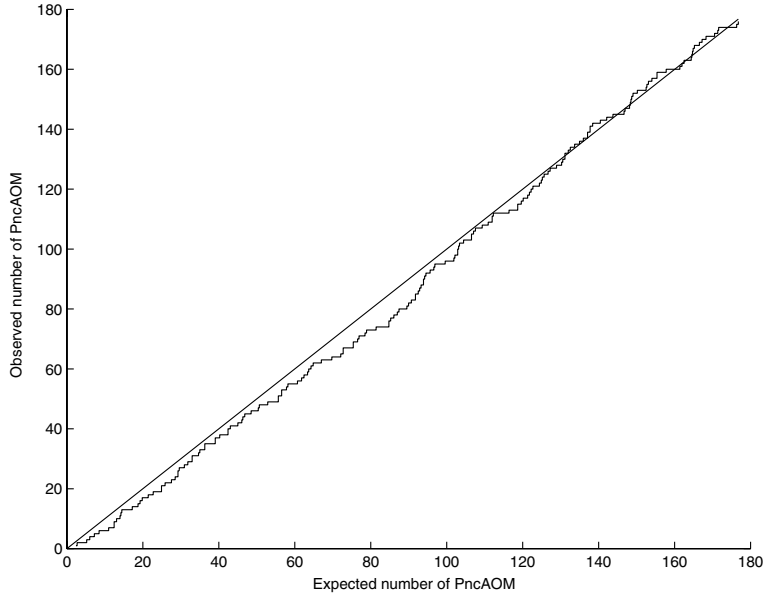

Fig. 10. TTT-plot: Observed (N_T) versus estimated ($\int \mu_t \, dt$) number of Pnc ear infections.

Andreev, 2000). When calculating cross-validation predictive probabilities, the observations of the ith child are erased from some chosen prediction time t^* until the end of the observation interval $(t^*, T]$. The predictive probabilities are then expectations of (14) given $\mathcal{F}^i_{t^*}$, and the possible sample paths of all processes of the ith individual are approximated by MCMC simulation.

In principle, it is possible to calculate the predictive number of ear infections either from some fixed prediction time onwards using the same data, or even for some replicated data of size n. Then the 'observations' for infection prevalence, viral infections and antibody measurements are simulated from processes with the estimated population parameters. In our case, the calculations become rather elaborate because of missing variables and the fact that all events are potentially recurrent. Apart from model assessment, such calculations have, however, other virtues also: by comparing relevant predictive probabilities, it is possible to monitor the influence of the timing of certain events of interest on the prediction (cf. Eerola, 1994). Such events of interest would in this study be, for example, viral infection when infected, or exceeding of a protective antibody level. However, as the other checks already convinced us about the stability of the present model, we do not report the preliminary predictive results here. In general, more methodological development is needed to investigate sensitivity and convergence of dynamic Bayesian models, although some papers have recently appeared in this area (e.g. Liechty & Roberts, 2001).

10. Discussion

We have presented a joint model for the dynamics of an individual infectious process which simultaneously updates information on infection, antibody response, carriage duration, and risk of recurrent disease which exists when infected.

Our modelling is related to the methodological literature of survival conditioned on a stochastic process where the interest is to obtain marginal results from the joint distribution of survival and the conditioning process. This involves averaging over the sample paths of the processes. Without simplifying assumptions it is hard to derive analytic results for such a case. In a series of papers, Manton, Woodbury & Yashin have presented models where a system failure is affected by a completely or partially unobserved stochastic process (e.g. Yashin & Manton, 1997). They provide backward smoothing equations which estimate the unobserved covariate paths of the past, and forward equations which update the parameter values in the light of new data. However, to obtain analytic results for the marginal survival distribution in simple terms of the conditioning stochastic process, the hazard of failure in their model needs to be a quadratic function of the conditioning process. Although suitable for some applications, this seems a restriction in many cases. In this respect, approximation of the joint distribution by MCMC simulation provides more freedom in modelling.

In the vast literature of HIV disease progression studies (e.g. Jewell & Kalbfleisch, 1992) the joint distribution of AIDS occurrence and a random marker process (CD4 versus CD8 counts after seroconversion) has been modelled to understand the predictive value of the evolution of the marker process for AIDS occurrence. The estimation of residual life length, given the marker process, corresponds to our prediction problem. The time origin usually refers to infection with the virus and failure indicates the actual onset of AIDS. Our model is a generalization of this design because the pattern of infection to disease may repeat randomly many times to the individuals.

Eerola (1994) used predictive probabilities to monitor the effect of one random event in a system of dependent event processes. Here this setting is generalized by letting one of the component processes (the antibody level) to vary continuously. In a system of event processes it is clear what a meaningful change is; it is the occurrence of the event of interest. In a continuously varying process this may not be as obvious. Some possibilities to limit the scope of estimation to causally meaningful changes (e.g. exceedings of a harmful level) is discussed in Eerola (1994). In many cases, the causal influence of such exceedings is not immediate but it is the cumulative stay which eventually leads to, for example, disease occurrence or its avoidance. In our application, a protective level of antibodies would be such a state of interest. However, the apparent effect of anti-PsaA antibodies was only marginal, and their protective action in this biological situation remained open.

One of the virtues of joint modelling and data augmentation is that it uses the data very effectively by combining the fragmentary pieces of information into the framework of a model. The modelling exercise is then not driven by the data collection pattern, as often in traditional statistical analysis, but rather by the postulated underlying biological mechanisms which vary continously in time. The role of the observed data becomes seemingly less important but in fact, as in our study, it is used to select the most probable pattern of the joint model. In our application, the major gain from modelling the complete system instead of observations was the construction of the infection episodes. This was the basis for the dynamic description of the infectious process. It is expected that the risk of AOM, and also its relationship to the explanatory factors (antibody level, viral infection), are estimated more accurately in the joint model because it mimics the true underlying sequence of events.

Although we have merely presented posterior estimates of some key parameters of the infectious process, it should be emphasized that once these ingredients have been estimated, it is possible to state any probabilistic question related to the infectious process, of course within the limits of the population it is representing. In this work, we have only touched upon the

question of prediction in complicated systems with recurrent, interdependent events. Future work will concentrate on this topic and on model assessment by predictive results.

Acknowledgements

The authors thank Jukka Corander and two anonymous referees for helpful comments, and the FinOM Study Group at KTL for the possibility to use the data. The work of D. Gasbarra has been supported by the Academy of Finland.

References

Arjas, E. & Andreev, A. (2000). Predictive inference, causal reasoning, and model assesment in nonparametric Bayesian analysis: a case study. *Lifetime Data Anal.* **6**, 187–204.

Arjas, E. & Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statist. Sinica* **4**, 505–524.

Auranen, K., Leino, T., Takala, A. & Arjas, E. (2000). Transmission of pneumococcal carriage in families: a latent Markov process model for binary longitudinal data. *J. Amer. Statist. Assoc.* **95**, 1044–1053.

Eerola, M. (1994). *Probabilistic causality in longitudinal studies*. Lecture Notes in Statistics Vol. 92. Springer, New York.

Ekholm, A., Jokinen, J. & Kilpi, T. (2002). Combining logistic regression and association modelling for longitudinal data on bacterial carriage. *Statist. Med.* **21**, 773–791.

Gelman, A. & Rubin, D. (1992). Inference from iterative simulations using multiple sequences. *Statist. Sci.* **7**, 457–511.

Geyer, C. & Møller, J. (1994). Simulation and likelihood inference for spatial processes. *Scand. J. Statist.* **21**, 359–373.

Gray, B., Converse III, G. & Dillon, H. Jr (1980). Epidemiologic studies of *Streptococcus Pneumoniae* in infants: acquisition, carriage, and infections during the first 24 months of life. *J. Infect. Dis.* **142**, 923–933.

Harper, M. (1999). Nasopharyngeal colonizations with pathogens causing otitis media: how does this information help us? *Pediatric Infect. Dis. J.* **18**, 1120–1124.

Jewell, N. & Kalbfleisch, J. (1992). Marker models in survival analysis and applications to issues associated with AIDS. In *AIDS Epidemiology: Methodological Issues* (eds N. P. Jewell, K. Dietz & V. T. Farewell), 211–230. Boston, Birkhäuser.

Kilpi, T., Herva, E., Kaijalainen, T., Syrjänen, R. & Takala, A. (2001). Bacteriology of acute otitis media in a cohort of Finnish children followed for the first two years of life. *Pediatric Infect. Dis. J.* **20**, 654–662.

Liechty, J. & Roberts, G. (2001). MCMC methods for switching diffusion models. *Biometrika* **88**, 299–315.

Rapola, S., Jäntti, V., Haikala, R., Syrjänen, R., Carlone, G., Sampson, J., Briles, D., Paton, J., Takala, A., Kilpi, T. & Käyhty, H. (2000). Antibodies to pneumococcal proteins PspA, PsaA and pneumolysin in children. Relation to pneumococcal nasopharyngeal carriage and otitis media. *J. Infect. Dis.* **182**, 1146–1152.

Ruuskanen, O. & Heikkinen, T. (1994). Viral–bacterial interaction in acute otitis media. *J. Pediatric Infect. Dis.* **13**, 1047–1049.

Smith, T., Lehmann, D., Montgomery, J., Graten, M., Riley, I. D. & Alpers, M. P. (1993). Acquisition and invasiveness of different serotypes of *Streptococcus pneumoniae* in young children. *Epidemiol. Infect.* **111**, 27–39.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82**, 528–550.

Vesa, S., Kleemola, M., Blomqvist, S., Takala, A., Kilpi, T. & Hovi, T. (2001). Epidemiology of documented viral respiratory infections and acute otitis media in a cohort followed from 2 to 23 months of age. *Pediatric Infect. Dis. J.* **20**, 574–580.

Yashin, A. & Manton, K. (1997). Effect of unobserved and partially observed covariate processes on system failure: a review of models and estimation strategies. *Statist. Sci.* **12**, 20–34.

Received December 2001, in final form January 2003

Mervi Eerola, Rolf Nevanlinna institute, PO Box 4, FIN-00014 University of Helsinki, Finland.
E-mail: mervi.eerola@rni.helsinki.fi

Appendix

The MCMC algorithm

Geyer & Møller (1994) give a 'birth and death' version of the Metropolis–Hastings algorithm for marked point processes. At every iteration, we add or remove randomly one point to or from the current configuration of infection times.

As before, we drop the individual superscript from the formulae. Denote the likelihood process of the antibodies X_t by

$$L_T^X(\rho^*) = \prod_{t \le T} \exp\left(-\frac{(X_t - X_{t-1} - \rho_t^*)^2}{2\sigma^2} \right) \tag{15}$$

where the product is over the time grid $t = 1, 2, \ldots, T$. The likelihood process for the AOM counting process N_t is

$$L_T^N(\mu) = \exp\left(\int_0^T \log(\mu_s) \mathrm{d}N_s + \int_0^T (1 - \mu_s) \mathrm{d}s. \right) \tag{16}$$

We describe in detail the updates for the individual infection process Y_t. This is the most non-standard part of the algorithm as the number of marked points in the infection configuration is a random variable itself. The other updates mainly rely on the full conditional densities described in section 6.

A1. Updating the infection process

We start with an individual infection pattern $\{\tau_1, \ldots, \tau_M\}$. To each point τ_k, associate a mark (ψ_k, d_k), where d_k is the duration of the infection and ψ_k the jump parameter of the antibody level at τ_k. We begin with an initial configuration $\{\tau_1^0, \ldots, \tau_{M_0}^0\}$ which is compatible with the Pnc carriage data $\{y_{s_j} : j = 1, \ldots, n\}$. This is obtained by having one infection for each j such that $y_{s_{j+1}} - y_{s_j} = +1$; by default we let the infection begin at $s_j + 1$, and end at $s_{j'} - 1$, where $j' = \min\{h : h > j \text{ and } y_{s_h} - y_{s_{h-1}} = -1\}$. This infection pattern has the minimal number of infection points compatible with the data. The Geyer & Møller update is as follows:

(i) With probability 1/2, propose to delete a point chosen uniformly at random among the points $\{(\tau_k, \psi_k, d_k)\}$ of the current marked point configuration,
Let τ^* be the infection time proposed for deletion; denote

$$\tau_{\text{next}} = T \wedge \min\{\tau_j : \tau_j > \tau^*\}$$
$$\tau_{\text{last}} = 0 \vee \max\{\tau_j : \tau_j < \tau^*\}$$
$$s' = \tau_{\text{last}} \vee \max\{s_h : \tau_{\text{last}} < s_h < \tau_{\text{next}} \text{ and } y_{s_{h+1}} - y_{s_h} = +1\}$$
$$s'' = \tau_{\text{next}} \wedge \min\{s_h : \tau_{\text{last}} < s_h < \tau_{\text{next}} \text{ and } y_{s_{h+1}} - y_{s_h} = -1\}$$

then if $s' > s''$ the move is immediately rejected. Otherwise, we resample the duration d_{last} of the infection at time τ_{last}, from the proposal distribution

$$q(d|\tau_{\text{last}}, \tau_{\text{next}}, s', s'') = \frac{f(d - c|\alpha, \nu)}{F(s'' - \tau_{\text{last}} - c|\alpha, \nu) - F(s' - \tau_{\text{last}} - c|\alpha, \nu)},$$

where f and F are the density and the c.d.f. of the prior Gamma(α, ν), respectively. For the moment we postpone the expression of the acceptance probability.

(ii) With probability 1/2, propose to add a marked point $(\tau_{M+1}, \psi_{M+1}, d_{M+1})$ to the infection pattern. The occurrence time τ_{M+1} of the proposed marked point is sampled from uniform distribution in [0, T].

Given the proposed infection time τ_{M+1} and the observations (y_{s_h}), denote

$$\tau_{\text{next}} = T \wedge \min\{\tau_j : \tau_j > \tau_{M+1}\},$$
$$\tau_{\text{last}} = 0 \vee \max\{\tau_j : \tau_j < \tau_{M+1}\},$$
$$s' = \tau_{\text{last}} \vee \max\{s_h : \tau_{\text{last}} < s_h < \tau_{M+1} \text{ and } y_{s_{h+1}} - y_{s_h} = +1\},$$
$$s'' = \tau_{M+1} \wedge \min\{s_h : \tau_{\text{last}} < s_h < \tau_{M+1} \text{ and } y_{s_{h+1}} - y_{s_h} = -1\},$$
$$s''' = \tau_{M+1} \vee \max\{s_h : \tau_{M+1} < s_h < \tau_{\text{next}} \text{ and } y_{s_{h+1}} - y_{s_h} = +1\},$$
$$s'''' = \tau_{\text{next}} \wedge \min\{s_h : \tau_{M+1} < s_h < \tau_{\text{next}} \text{ and } y_{s_{h+1}} - y_{s_h} = -1\}.$$

The proposed pattern is compatible with the data iff

$$s' < s'' \leq s''' < s'''',$$
$$s'' - \tau_{\text{last}} > c, \text{ and } s'''' - \tau_{M+1} > c,$$

as the durations are assumed to last at least as long as the antibody response delay $c = 14$ days. If one of these conditions is not fullfilled, the proposed configuration is immediately rejected.

Otherwise, we continue, and propose new durations d_{M+1} and d_{last} for the respective infections at τ_{M+1} and τ_{last}.

The proposal for d_{M+1} is the truncated density

$$q(d|\tau_{M+1}, \tau_{\text{next}}, s''', s'''') = \frac{f(d - c|\alpha, v)}{F(s'''' - \tau_{M+1} - c|\alpha, v) - F(s''' - \tau_{M+1} - c|\alpha, v)}.$$

Analogously, the duration starting at time τ_{last} is sampled from the truncated density

$$q(d|\tau_{M+1}, \tau_{\text{next}}, s', s'') = \frac{f(d - c|\alpha, v)}{F(s'' - \tau_{M+1} - c|\alpha, v) - F(s' - \tau_{M+1} - c|\alpha, v)}.$$

(iii) To have better mixing, the proposal density $q(\psi; \tau_{M+1}, X)$ of the mark ψ_{M+1} was allowed to depend on the state of the process X, that is, on τ_k and the increment of X in $[\tau_{M+1}, \tau_{M+1} + c]$. It is given by a truncated normal distribution restricted to the positive half line, with variance

$$\tilde{\sigma}^2 = \frac{1}{(b_0^2/a_0 + \min(1, (X_{\tau_{M+1}})^{-2}) \cdot c^2/\sigma^2)},$$

and mean

$$\tilde{m} = \tilde{\sigma}^2 \left(b_0 + \frac{\min(1, (X_{\tau_{M+1}})^{-1}) \times (X_{\tau_{M+1}+c} - X_{\tau_{M+1}} + c\rho)}{\sigma^2} \right).$$

Following Geyer & Møller, the new point $(\tau_{M+1}, \psi_{M+1}, d_{M+1})$ is accepted with probability

$$1 \wedge \left(\frac{L_T^X(\rho^{*\text{new}}) L_T^N(\mu^{\text{new}})}{L_T^X(\rho^{*\text{old}}) L_T^N(\mu^{\text{old}})} \frac{p(\psi_{M+1}; a_0, b_0)}{(M+1) q(\psi_{M+1}; \tau_{M+1}, X)} \right),$$

where $(\rho_t^{\text{new}}, \mu_t^{\text{new}})$ and $(\rho_t^{\text{old}}, \mu_t^{\text{old}})$ are the parameter processes of proposed and old configurations in (15) and (16) and $p(\psi_{M+1}; a_0, b_0)$ is the prior of ψ. To make the move reversible, a uniformly chosen marked point (τ_m, ψ_m, d_m) is deleted with probability

$$1 \wedge \left(\frac{L_T^X(\rho^{*\text{new}}) L_T^N(\mu^{\text{new}})}{L_T^X(\rho^{*\text{old}}) L_T^N(\mu^{\text{old}})} \frac{(M+1) q(\psi_{M+1}; \tau_{M+1}, X)}{p(\psi_{M+1}; a_0, b_0)} \right).$$

A2. Updating the antibody process

At the observation times t_k, the antibody process is given the observed values, i.e. $X_{t_k} = x_{t_k}$.

Otherwise the conditional distribution of $[X_t | Y, X_s, s \neq t]$ is proportional to

$$\exp\left[-\frac{1}{\sigma^2 \Delta t}\left(X_t - \frac{X_{t-1} + X_{t+1} + (\rho^*_{t-1} - \rho^*_t)\Delta t}{2}\right)^2\right]$$

which is used as a proposal distribution.

The Hastings ratio is

$$\frac{L^X_T(\rho^{*\text{new}})}{L^X_T(\rho^{*\text{old}})}\frac{L^N_T(\mu^{\text{new}})}{L^N_T(\mu^{\text{old}})}.$$

A3. Updating the regression and the baseline processes

We describe the MCMC updates for the regression process α_t; the same moves are used to update $\log \mu_0(t)$ and $\log \lambda_{01}(t)$.

The process (α_t) is piecewise constant with change points $(R_1, \ldots, R_M, R_{M+1} = T)$, and levels $(\alpha_1, \ldots, \alpha_M, \alpha_{M+1})$. The prior dynamics was given in section 5.

In the Geyer & Møller algorithm: with probability 1/2, propose a new change point R^* from the normalized driving measure $\eta(\mathrm{d}t)/\eta([0, T])$. For $R_{i-1} < R^* < R_i$, the corresponding new levels in the proposal configuration are

$$\alpha^*_i = \frac{U \cdot \alpha_i \cdot (R_i - R_{i-1})}{(R^* - R_{i-1})}, \quad \text{and} \quad \alpha^*_{i+1} = \frac{(1-U) \cdot \alpha_i \cdot (R_i - R_{i-1})}{(R_i - R^*)},$$

where U is an independent random variable uniform in [0, 1]. The Jacobian of the transformation $\mathcal{T}:(\alpha_i, U) \to (\alpha^*_i, \alpha^*_{i+1})$ is

$$J(\mathcal{T}) = \alpha_i \cdot \frac{(R_i - R_{i-1})^2}{(R^* - R_{i-1})(R_i R^*)}.$$

Note that this move preserves the area $\int_0^T \alpha(t)\mathrm{d}t$.

With probability 1/2, propose to delete a point R_i with i chosen uniformly at random in $\{1, \ldots, M\}$. In this case, to make the algorithm reversible, the levels are updated in the following way: delete α_{i+1}, relabel the points accordingly, and set

$$\alpha^*_i = \frac{\alpha_{i+1}(R_{i+1} - R_i) + \alpha_i(R_i - R_{i-1})}{R_{i+1} - R_{i-1}}.$$

When we add a marked point, the Hastings ratio is

$$\frac{L^N_T(\mu^{\text{new}})}{L^N_T(\mu^{\text{old}})}\frac{\eta([0, T]) \cdot J(\mathcal{T})}{(M+1)}\frac{\text{Prior}(\alpha^{\text{new}} | S^{\text{new}})}{\text{Prior}(\alpha^{\text{old}} | S^{\text{old}})}.$$

and the reciprocal gives the Hastings ratio for a deletion.

To complete the algorithm, we add a move where update in turn each level α_i, $i = 1, \ldots, M$ conditionally on the change points $R_1, \ldots, R_{M+1}$, by sampling from the random walk proposal

$$\alpha^{\text{new}}_i = \mathcal{N}(\alpha^{\text{old}}_i, \varepsilon^2), \quad \varepsilon = 0.05.$$