Hierarchical Nonparametric Bayesian Models for the Force of Infection for Mumps and Rubella

Z. Shkedy\(^1\), M. Aerts\(^1\), G. Molenberghs\(^1\), Ph. Beutels\(^2\) and P. Van Damme\(^2\)

\(^1\) Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus - gebouw D, B 3590 Diepenbeek, Belgium

\(^2\) University of Antwerp, Epidemiology and Community Medicine, Center for Evaluation of Vaccination, B 2610 Antwerp, Belgium

**Summary**

The force of infection and average age at infection are of primary interest in the analysis of infectious diseases. Parametric approaches to estimate the force of infection from serological data are based on generalized linear models (GLM) and nonlinear models for the observed prevalence. Isotonic regression and monotone local polynomials can be used as alternative nonparametric methods. Within the Bayesian framework, hierarchical GLM and nonlinear models can be used to estimate the force of infection parametrically. This paper introduces nonparametric Bayesian hierarchical models for the force of infection. A prior distribution (product Beta) is specified in a way that the posterior mean of the prevalence satisfies the monotonicity constraints and leads to a non-negative force of infection. In this approach the method can be viewed as isotonic Bayesian regression.

*Correspondence to: Ziv Shkedy, Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus - gebouw D, B–3590 Diepenbeek, Belgium, email: ziv.shkedy@luc.ac.be
Using examples of cross sectional serological samples of Rubella and Mumps, this paper shows the applicability of the method and its advantage in terms of flexibility in estimating the prevalence, the force of infection and other epidemiological parameters related to the disease.

Keywords: Constrained Parameters; Force of Infection; Hierarchical Bayesian Models; Infectious Diseases; Isotonic Regression; Prevalence.

1. Introduction

The transmission process of infectious diseases is governed by the force of infection, which is the rate at which the susceptible individuals acquire an infection (Grenfell and Anderson [1] and Anderson and May [2]). Under the assumptions of life-long immunity and that the disease does not effect mortality and that death rate is zero up to life expectancy $L$, the change in the susceptible class with respect to the host age $a$ is described by the differential equation

$$\frac{d}{da} \pi(a) = -\ell(a)\pi(a),$$

where $\pi(a)$ is the probability to be susceptible at age $a$, the prevalence is $\theta(a) = 1 - \pi(a)$ and $\ell(a)$ is the force of infection (or the hazard function), $\ell(a) = \theta'(a)/(1 - \theta(a))$. Equation (1) represents the change in the proportion of susceptibles in the population when the disease is in the steady state and it can be shown that $\theta(a) = 1 - \exp(-\int_0^a \ell(s)ds)$.

The prevalence can be estimated from a cross-sectional seroprevalence sample, which contains the information about the current disease status (infected yes/no) and the age of the individual at test (Diamond and Mcdonald [3]). Figure 1 shows two cross-sectional seroprevalence samples of rubella and mumps that were recently discussed by Farrington et al. [4] that will be used to illustrate the methods in this paper.
Frequentist parametric models were discussed by Muench [5], Griffiths [6] and Gernfell and Anderson [1], who used GLM for binary responses with a log link function to estimate the prevalence and the force of infection. In these models the so-called linear predictor is given by

$$\eta(a) = \sum_{i=0}^{k} \beta_i a^i$$

and the force of infection is simply the first derivative of the linear predictor. However, the main difficulty of this approach is that it can lead to non-monotone estimates for the prevalence and therefore to negative values of the force of infection. Several other parametric models within the framework of GLM for binary response were suggested by Becker [7], Diamond and Mcdonald [3] and Kieding [8]. Farrington [9] and Farrington et al. [4] proposed nonlinear models for the force of infection. In Farrington [9] and Edmunds et al. [10] the model for the force of infection is defined by

$$\ell(a) = (\alpha_1 a - \alpha_3)e^{-\alpha_2} + \alpha_3. \quad (2)$$

In order to ensure that the force of infection satisfies $$\ell(a_i) \geq 0, i = 1, 2, \ldots, n$$, Farrington [9] constrained the parameter space to be nonnegative ($$\alpha_j \geq 0, j = 1, 2, 3$$). This model is further discussed in detail in Section 3.1.

Within the framework of fully nonparametric Bayesian modeling, the problem is to estimate $$\theta_1, \theta_2, \ldots, \theta_n$$ under the order restriction $$\theta_1 \leq \theta_2 \leq \cdots \leq \theta_n$$ where $$\theta_i = \theta(a_i)$$. The prevalence is assumed to be an isotonic nonparametric function satisfying $$0 \leq \theta_i \leq 1$$. Here, we estimate the posterior distribution $$P(\theta_1, \theta_2, \ldots, \theta_n | y)$$, where $$y = (y_1, y_2, \ldots, y_n)$$ and $$y_i$$ is the number of infected individuals at age $$a_i$$. The $$n$$ dimensional parameter vector is constrained to lie in a subset $$S^n$$ of $$\mathbb{R}^n$$. The constrained set $$S^n$$ is determined by the order restrictions among the components of $$\theta$$, $$\theta = (\theta_1, \theta_2, \ldots, \theta_n)$$. In this case it is natural to incorporate the constraints into the specification of the prior distribution, $$P(\theta)$$. In the context of bioassay modeling,
Gelfand and Kuo [11] showed that the constrained posterior distribution has the same form as the unconstrained posterior distribution restricted to the constrained set. This implies that if $P(\theta)$ is a product-Beta distribution, and the likelihood $P(y|\theta)$ is binomial, then the posterior distribution $P(\theta_i|y_i, \theta_j, i \neq j)$ is a Beta distribution restricted to the interval $[\theta_{i-1}, \theta_{i+1}]$. In a parametric Bayesian framework $\theta$ has a parametric form, $\theta(a, \alpha)$, where $\alpha$ is a parameter vector. In this case $\theta$ has a deterministic relationship with the predictor $a$ and one may need to constrain the parameter space of the prior distribution $P(\alpha)$ in order to achieve monotonicity of the posterior distribution $P(y|\theta)P(\theta|\alpha)P(\alpha)$, for example, constrain the parameters to be nonnegative as done by Farrington [9] in a frequentist setting.

This paper is organized as follows. In Section 2, we present an exploratory data analysis based on isotonic regression and local polynomial estimation for the prevalence and the force of infection. In Section 3, non-linear and generalized linear hierarchical Bayesian models are implemented. Model selection is based on the Deviance Information Criterion, (DIC). In Section 4 we relax the parametric assumption about the mean structure of the prevalence and discuss the nonparametric Bayesian model. The methods are illustrated using the datasets of Rubella and Mumps.

2. Exploratory Data Analysis

Kieding [12] proposed a nonparametric estimation procedure for both the prevalence and the force of infection. Keiding’s [12] procedure consists of two steps in which in a first step the prevalence is estimated using isotonic regression (Barlow et al., [13] and Robertson et al. [14]). The isotonic estimator is calculated by applying the pool adjacent violator algorithm (PAV) to the observed prevalence resulting in a monotone step function as an estimator for the
prevalence. In a second step, the force of infection, which is assumed to be a smooth function, can be derived from the prevalence using a kernel smoother. Shkedy et al. [15] showed that for GLM for a binary response, the force of infection has the general form \( \ell(a) = \eta'(a)\delta(\eta(a)) \), where \( \eta(a) \) is the linear predictor and the form of \( \delta(.) \) is determined by the link function used in the model. Using this general form for the force of infection, Shkedy et al. [15] proposed to estimate the prevalence and the force of infection with local polynomials. In this approach, one first smooths the data with a local polynomial and then monotonize (if necessary) the estimate in order to ensure a nonnegative force of infection. Both smoothing methods need a prior choice of the bandwidth. Keiding [12] selected the bandwidth by visual inspection whereas Shkedy et al. [15] derived a data driven optimal bandwidth minimizing the asymptotic local mean square error of the force of infection. Figure 2 shows both the local polynomial and the isotonic regression estimates for \( \pi(a) \) and \( \ell(a) \). For rubella (see Figure 2, panel b), the estimated force of infection, estimated by the local quadratic model, rises steeply to a peak at age 7-8 followed by a steady decrease to zero at older age groups. The two methods result in somewhat different patterns. The force of infection estimated by the kernel estimate predicts a secondary peak at age 24 and a third peak at age 40. Panel d in Figure 2 reveals the same patterns for mumps. We note that the second peak at age 10 estimated by the local polynomial is smoothed out by the kernel smoother which also predicts a third peak at age 33.

**Figure 2, About Here.**
3. Hierarchical Bayesian Models for the Force of Infection

3.1. Non-linear Hierarchical Model

The model in (2) assumes that the force of infection is zero at birth ($\ell(0) = 0$) and then rises to a peak in a linear fashion followed by an exponential decrease. The peak is reached at an age corresponding to the maximum contact rate of susceptibles with infectious individuals. The parameter $\alpha_3$ is called the long term residual value of the force of infection. If $\alpha_3 = 0$, then the force of infection decreases to 0 as $a$ tends to infinity. Integrating $\ell(a)$ results in a nonlinear model

$$\pi(a) = 1 - \exp \left\{ \frac{\alpha_1}{\alpha_2}ae^{-\alpha_2a} + \frac{1}{\alpha_2} \left[ \frac{\alpha_1}{\alpha_2} - \alpha_3 \right] \left[ e^{-\alpha_2a} - 1 \right] - \alpha_3a \right\}. \quad (3)$$

In what follows we refer to (3) as the exponential model. The average age at infection, the mean of the distribution of the age of infection, is given by $A = \int_0^L (1 - \pi(x))dx$, where $L$ is life expectancy. Following Farrington [9] we assume that $L = 75$. In case that the data are observed up to a certain age $U$, $U \leq L$, the average age at infection is given by

$$A = \int_0^U (1 - \pi(x))dx + f(L - U). \quad (4)$$

Here, $f$ is the fraction of individuals that remain uninfected which can be estimated from the data by $f = 1 - \pi(U)$. Farrington [9] estimated unrestricted models for measles, mumps and rubella based on (3) and performed sensitivity analysis for $f$ by estimating the model in (3) conditional on several values for $f$. In these analyses the parameter $\alpha_1$ is no longer a free parameter but can be calculated conditional on the values of $\alpha_2$, $\alpha_3$ and $f$.

In the present study we use hierarchical nonlinear models to estimate the parameters in the exponential model (3). Independent binomial distributions are assumed for the numbers of
infected individuals at age $a_i$

$$y_i \sim \text{Bin}(n_i, \pi_i), \quad \text{for } i = 1, 2, \ldots, n,$$

where $n_i$ is the sample size at age $a_i$. The constraints on the parameter space can be incorporated in the hierarchical model by assuming truncated normal distributions for the components of $\alpha$, $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, in $\pi_i = \pi(a_i, \alpha)$,

$$\alpha_j \sim \text{truncated } N(\mu_j, \tau_j) \quad j = 1, 2, 3.$$

Here, the normal prior distribution is left truncated at 0 to ensure that $\ell(a) \geq 0$. The joint posterior distribution for $\alpha$ can be derived by combining the likelihood and the prior model as

$$P(\alpha|y) \propto n \prod_{i=1}^{n} \text{Bin}(y_i|n_i, \pi(a_i, \alpha)) \prod_{j=1}^{3} \frac{1}{\tau_j} \exp\left(\frac{1}{2\tau_j^2}(\alpha_j - \mu_j)^2\right).$$

The full conditional distribution of $\alpha_i$, derived from (6), is given by

$$P(\alpha_i|\alpha_j, \alpha_k, k, j \neq i) \propto \frac{1}{\tau_i} \exp\left(\frac{1}{2\tau_i^2}(\alpha_i - \mu_i)^2\right) n \prod_{i=1}^{n} \text{Bin}(y_i|n_i, \pi(a_i, \alpha)),$$

which cannot be simplified further. To complete the specification of the probability model we assume flat hyperprior distributions at the third level of the model, i.e. $\mu_j \sim N(0, 10000)$ and $\tau_j^{-2} \sim \text{gamma}(1000, 1000)$.

### 3.2. Hierarchical Log-logistic Model

The exploratory analysis from Section 2 indicates that the force of infection rises to a peak and drops down thereafter. Therefore we can conclude that the time spent in the susceptible class is not an outcome of neither an exponential nor a Weibull distribution since these distributions have a constant and a monotone force of infection respectively. In contrast, the log-logistic distribution offers a wide range of curve shapes for the hazard function, which is more capable
to capture the common pattern revealed in Figure 2 (although, similar to model (3), the secondary peaks will be smoothed out). Under the assumption that the time spent in the susceptible class follows a log-logistic distribution, the probability to become infected before age $a$ is given by

$$\pi(a) = \frac{\beta a^\alpha}{1 + \beta a^\alpha}, \quad \alpha, \beta > 0,$$

and the force of infection by

$$\ell(a) = \frac{\alpha \beta a^{\alpha-1}}{1 + \beta a^\alpha}.$$

The log-logistic model can be fitted as a GLM with $\log(a)$ as a predictor and a logit link function. This leads to a Bayesian logistic regression model (Gilks et al. [16] and Gelman et al. [17]) of $y$ with covariate $\log(a)$. We specify the same likelihood as in (5) with linear predictor given by

$$\logit(\pi(a)) = \alpha_2 + \alpha_1 \log(a),$$

where $\alpha_2 = \log(\beta)$. For the prior model of $\alpha_1$ ($\alpha_1 = \alpha$), we specify $\alpha_1 \sim \text{truncated } N(\mu_1, \tau_1)$. We constrain $\beta$ to be positive by specifying $\alpha_2 \sim N(\mu_2, \tau_2)$. The full conditional distribution of $\alpha_1$ is

$$P(\alpha_1 | \alpha_2) \propto \frac{1}{\tau_1} \exp \left( \frac{1}{2\tau_1^2}(\alpha_1 - \mu_1)^2 \right) \prod_{j=1}^n \text{Bin}(y_i | n_i, \pi(a_i, \alpha_1, \alpha_2)).$$

(10)

The full conditional distribution for $\alpha_2$ can be derived in the same way. The same flat hyperprior distributions as in the previous section are assumed for the hyperparameters.

3.3. Model Selection

Within the Bayesian framework, the unknown parameters are estimated by the posterior mean. However, since the full conditional distributions in (7) and (10) do not have a closed analytical form, we cannot evaluate it directly. We can approximate it using Markov Chain Monte
Carlo (MCMC) methods (Gilks et al. [16]) and generate samples form the full conditional distributions using the Gibbs sampler. The sample averages are taken as the posterior means of the parameters of interest.

A model selection procedure is needed in order to compare between the models mentioned above and to select the best model. Goodness-to-fit and complexity of the models were assessed using the deviance information criterion (DIC) as proposed by Spiegelhalter et al. ([18], [19]) and recently used by Erkanli et al. [20], Rahmann et al. [21] and Gelfand et al. [22] for model selection within the Bayesian framework. Spiegelhalter et al. [18] and [19] suggested to measure the effective number of parameters (the complexity) in the model by the difference between the posterior expectation of the deviance and the deviance evaluated at the posterior expectation of \( \pi \), that is

\[
P_D = E_{\pi|y}(D) - D(E_{\pi|y}(\pi)) = \bar{D} - D(\bar{\pi}),
\]

with deviance given by \( D(\pi) = -2 \log P(y|\pi) + 2 \log(f(y)) \). The second term in the deviance is a standardizing factor which does not depend on \( \pi \); we use -2 log likelihood of the saturated model. Hence, for the models discussed above the binomial deviance is given by

\[
D(\pi) = 2 \sum_i \left( y_i \log \frac{y_i}{n_i \pi_i} + (y_i - n_i) \log \frac{1 - y_i}{1 - \pi_i} \right).
\]

In practice, \( D(\pi) \) and \( \pi \) can be monitored during the MCMC run, \( \bar{D} \) is the sample mean of \( D(\pi) \) while \( D(\bar{\pi}) \) is the deviance evaluated at the posterior mean. For model selection, Spiegelhalter et al. [18] and [19] suggested to use the Deviance Information Criterion (DIC):

\[
DIC = \bar{D} + P_D = D(\bar{\pi}) + 2P_D.
\]

Smaller values of \( DIC \) indicate a better fitting model.
3.4. Application to the Data

Table I presents the deviance summaries of the data and Figure 3 shows the fitted models for both the prevalence and the force of infection. Starting with rubella, the first model that was fitted assumes that $\alpha_3 = 0$ in the exponential model in (3). For this model the posterior deviance is 64.7 and $P_D = 2.02$, slightly higher than the “true” number of parameters. For the exponential model with $\alpha_3 > 0$, $\bar{D} = 61.13$ and $P_D = 3.00$. The $DIC$ of this model is 64.13, smaller than the $DIC$ of the first model (66.72) indicating that among the exponential models the second one is to be preferred. However, the log-logistic model with $DIC = 61.28$ has the best goodness-to-fit. For mumps, the model with the lowest value of $DIC$ is the exponential model with $\alpha_3 = 0$ (65.13). Posterior means for the parameters are shown in Table II. Figure 3 shows that there is a substantial difference between the models at the age for which the force of infection reaches its peak and in the level of the force of infection at older age groups. Furthermore, for rubella, the posterior mean of the average age at infection for the exponential model with $\alpha_3 > 0$ is 10.16 and the posterior mean for $f^*$ is 0.04. When $\alpha_3$ is not included in the model, the average age of infection increases to 11.11 and $f^*$ to 0.07. The posterior mean of the average age at infection obtained from the log-logistic model is 9.86. For mumps, the effect of $\alpha_3$ on $f^*$ is less substantial, the reason for that is the small value of $\alpha_3$ that was estimated in the second model (0.008). The smallest value of $f^*$ is obtained for the log-logistic model (0.0007) with average age at infection equal to 5.053.

TABLE 1, TABLE 2 AND FIGURE 3 ABOUT HERE.
4. Hierarchical Nonparametric Model

4.1. Hierarchical Beta/Binomial Model

In the previous section the prevalence was assumed to have a parametric form \( \pi(a, \alpha) \), and monotonicity was achieved by constraining the parameter space of \( \alpha \). In this section, we assume that \( \pi \) is a right-continuous nondecreasing function defined on \([0, \delta]\), \( \pi_n \leq \delta \leq 1, \delta = 1 - f \). We do not assume any deterministic relationship between \( \pi_i \) and \( a_i \) but instead we specify a probabilistic model for \( \pi_i \) at each distinct level of \( a_i \). Since the data are binomial, it is natural to use the product-beta prior (Gelfand and Kuo [11]) for \( \pi \), since it is a conjugate prior for the binomial likelihood and ensures that the posterior distribution of \( \pi | y \) is also a beta distribution. A product-beta prior has the form of

\[
P_B(\pi | \alpha, \beta) \propto \prod_{i=1}^{n} (\pi_i)^{\alpha_i - 1} (1 - \pi_i)^{\beta_i - 1}, \quad (\alpha_i > 0, \beta_i > 0),
\]

where \( \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n) \) and \( \beta = (\beta_1, \beta_2, \ldots, \beta_n) \). For the unconstrained case, combining the binomial likelihood and the product-beta prior, leads to the posterior distribution

\[
P(\pi | y, \alpha, \beta) \propto \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \prod_{i=1}^{n} \pi_i^{\alpha_i - 1} (1 - \pi_i)^{\beta_i - 1} = \prod_{i=1}^{n} \pi_i^{y_i + \alpha_i - 1} (1 - \pi_i)^{n_i - y_i + \beta_i - 1},
\]

which is Beta(\(y_i + \alpha_i, n_i - y_i + \beta_i\)). The problem is to estimate \( \pi \) under the order restrictions, \( \pi_1 \leq \pi_2 \leq \cdots \leq \pi_n \). Thus, the n dimensional parameter vector is constrained to lie in a subset \( S^n \) of \( \mathbb{R}^n \). The constrained set \( S^n \) is determined by the order among the components of \( \pi \). In this case it is natural to incorporate the constraints into the specification of the prior distribution. Gelfand, Smith and Lee [23] show that the posterior distribution of \( \pi \) given the constraints is the unconstrained posterior distribution normalized such that

\[
P(\pi | y) \propto \int_{S^n} \frac{P(y | \pi) P(\pi | \alpha, \beta)}{P(y | \pi) P(\pi | \alpha, \beta) d\pi}, \quad \pi \in S^n.
\]
Let $S^n_j(\pi_j, j \neq i)$ be a cross section of $S^n$ defined by the constraints for the component $\pi_i$ at a specified set of $\pi_j, j \neq i$. In our setting, $S^n_j(\pi_j, j \neq i)$ is the interval $[\pi_{i-1}, \pi_{i+1}]$. It follows from (16) that the posterior distribution for $\pi_i$ is given by

$$
P(\pi_i | y, \alpha, \beta, \pi_{-i}) \propto P(y | \pi)P(\pi | \alpha, \beta) \quad \pi_i \in S^n_j(\pi_j, j \neq i),$$

$$0, \quad \pi_i \notin S^n_j(\pi_j, j \neq i).$$

(17)

Here, $\pi_{-i} = (\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n)$. Hence, when the likelihood and the prior distribution are combined, the posterior conditional distribution of $\pi_i | y, \alpha, \beta, \pi_{-i}$ is the standard posterior distribution restricted to $S^n_j(\pi_j, j \neq i)$, that is $\text{Beta}(y_i + \alpha_i, n_i - y_i + \beta_i)$ restricted to the interval $[\pi_{i-1}, \pi_{i+1}]$ (Gelfand and Kuo [11]). This means that during the MCMC simulation the sampling from the full conditional distribution can be reduced to interval restricted sampling from the standard posterior distribution (Gelfand, Smith and Lee [23]).

The hierarchical model we consider is given by

$$
y_i \sim \text{Bin}(n_i, \pi_i) \quad \text{likelihood}$$

$$
\pi_i \sim \text{Beta}(\alpha_i, \beta_i)I(\pi_{i-1}, \pi_{i+1}) \quad \text{prior},$$

(18)

where $I(\pi_{i-1}, \pi_{i+1})$ an indicator variable which takes the value of 1 if $\pi_{i-1} \leq \pi_i \leq \pi_{i+1}$ and zero elsewhere. In order to complete the specification of the hierarchical model in (18) we need to specify a hyperprior distributions for $\alpha$ and $\beta$. Note that the special case that $\alpha_i = \beta_i = 1$ for $i = 1, \ldots, n$ implies that the prior distribution of the prevalence at the $i$'th age group, conditional on $\pi_{i-1}$ and $\pi_{i+1}$, is a uniform distribution over the interval $[\pi_{i-1}, \pi_{i+1}]$. 
$$\pi_i|\pi_{i-1}, \pi_{i+1} \sim \text{Uniform}(\pi_{i-1}, \pi_{i+1})$$. However, there is no reason to fix $\alpha$ and $\beta$ to be equal to 1, there is no clear way how to choose the hyperprior distribution for the components in $\alpha$ and $\beta$ either. For the analysis presented below we specify non informative distributions for the hyperparameters by specifying a left truncated (at zero) normal distribution with variance equal to 1000 for each one of the components in $\alpha$ and $\beta$ at the third stage of the hierarchical model.

Once the prevalence values are obtained, the problem of estimating the force of infection becomes straightforward. Let $\pi^{(k)}$ be the constrained value of $\pi$, obtained in the $k$’th iteration of the MCMC simulation. The force of infection $\ell^{(k)}(a)$ can be estimated by

$$\hat{\ell}^{(k)}(a) = \pi^{(k)}(a)/(1 - \hat{\pi}^{(k)}(a))$$. However, since we assume that the force of infection is a smooth function, we smooth $\ell^{(k)}(a)$ with a twice successively third order moving average (Diggle [26]), i.e. $\ell_S^{(k)}(a) = S\ell^{(k)}(a)$ where $\ell_S(a)$ is the smoothed force of infection and $S$ is the smoothing matrix. The posterior mean of $\ell_S(a)$ is simply $\sum_{k=1}^K \ell_S^{(k)}(a)/K$ where $K$ is the number of MCMC iterations.

The fraction of uninfected individuals can be used in this model to specify the distribution of $\pi(U)$. If our prior assumption is that $f = 0$, then $\pi(U) \sim \text{Beta}(\alpha_n, \beta_n)I(\pi_{n-1}, 1)$, where $I(\pi_{n-1}, 1)$, is an indicator that takes the value of 1 if $\pi_{n-1} \leq \pi_n \leq 1$ and 0 otherwise. In case that we use the prior knowledge that $f > 0$, say $f = f^*$, then we can truncate the distribution of $\pi(U)$ at the right side with $1 - f^*$, $\pi(U) \sim \text{Beta}(\alpha_n, \beta_n)I(\pi_{U-1}, 1 - f^*)$.

4.2. Application to the Data

The posterior means for the prevalence and force of infection are shown in Figure 4 (rubella) and Figure 5 (mumps). For rubella, the nonparametric models indicate essentially the same
patterns, although the secondary peak at age 23 is less substantial in the beta-binomial model. In addition, from age 30 onwards, the beta-binomial model predicts a higher force of infection. For mumps, the secondary peak at age 20 was smoothed out by the beta-binomial model. Similar to rubella, the beta-binomial model predicts higher values for the force of infection at the first peak, compared to the parametric model. This can be seen in Figure 6 which presents the density estimates for the posterior distribution of the force of infection between age 3.5 and 6.5. Note that the exponential and the beta-binomial models for the force of infection reach a peak at age 4.5 and 5.5 respectively. The beta-binomial model predicts higher values for the force of infection at the ages, 0.36 and 0.29 for the beta-binomial and the exponential models respectively.

The value of $f$ has a substantial influence on the posterior mean of the average age at infection. We fitted the beta-binomial model with several values of $f$. That is, we truncated the distribution of $\pi(U)$ at the right hand side with $1 - f$, $\pi(U) \sim \text{Beta}(\alpha_n, \beta_n)I(\pi_U - 1, 1 - f)$. Table III presents the results and shows that the posterior mean of $A$ increases with $f$. This can be seen in Figure 8 which shows the 95% credible intervals for the average age at infection. This pattern was observed by Farrington [9] for the estimated conditional models (see Farrington [9], Table 3). Note that in the second column in Table III, $\bar{f} = 1 - \bar{\pi}(U)$, is the posterior mean for $f$. Figure 7 shows the estimated forces of infection for several values of $f$. Note that substantial differences are observed from age 30 and onwards. The force of infection increases with higher values of $f$.

FIGURE 4, FIGURE 5, FIGURE 6, ABOUT HERE.
5. Discussion

The age-dependent force of infection is a basic concept in any epidemiological model for infectious disease. Furthermore, the average age at infection and the basic reproduction number, $R_0$, depend on the model for the force of infection. In this study we model the prevalence within the framework of hierarchical Bayesian models in order to investigate the posterior distribution of $\ell(x)$ and $A$. The parametric models are restrictive since they can estimate only a single peak model for the force of infection. However, the beta-binomial model suggests secondary peaks which may be important from an epidemiological point of view. Furthermore, we have shown that compared to the parametric models, the beta-binomial models predict higher values for the force of infection at its maximum.

Within the framework of frequentist nonparametric estimation, the problem of estimation under order restrictions was addressed by applying the PAV algorithm to the nonparametric estimator of the prevalence. For the beta-binomial model, monotonicity was achieved by choosing a truncated product-Beta for the prior model in the hierarchical beta-binomial model. Both models estimate a nondecreasing prevalence and therefore lead to a nonnegative force of infection, as required. The beta-binomial model is highly sensitive for the values of $f$. It is necessary to fit the model with several values of $f$ in order to investigate its influence on the posterior mean of $A$ and $\ell(x)$. We specified a product-beta as a prior distribution for the hierarchical nonparametric model. An order Dirichlet distribution as discussed, in the context of binary response, by Ramsey [24], Gelfand and Kuo [11] and Qian et al. [25] can be used as well. This issue is a subject for future investigation.

TABLE 3, FIGURE 7, FIGURE 8 AND FIGURE 6, ABOUT HERE.
REFERENCES


Chapman and Hall, London.


Table I. Deviance and goodness-of-fit summaries for the parametric Bayesian models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Rubella</th>
<th>Mumps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{D}$</td>
<td>$D(\pi)$</td>
</tr>
<tr>
<td>Exponential</td>
<td>64.70</td>
<td>62.70</td>
</tr>
<tr>
<td>($\alpha_3 = 0$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponential</td>
<td>61.13</td>
<td>58.13</td>
</tr>
<tr>
<td>($\alpha_3 &gt; 0$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log-logistic</td>
<td>58.99</td>
<td>59.70</td>
</tr>
</tbody>
</table>
Table II. Posterior means for the parameters. Note that the $\alpha$ parameters of the log logistic model and the exponential model are not comparable. The exponential model with $\alpha_3 = 0$ is the model proposed by Farrington (1990) with the assumption that $\ell(a) = \alpha_1 a \exp(-\alpha_2 a)$.

<table>
<thead>
<tr>
<th></th>
<th>Rubella</th>
<th>Mumps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exponential</td>
<td>Exponential</td>
</tr>
<tr>
<td>$\alpha_3 = 0$</td>
<td>$\alpha_3 &gt; 0$</td>
<td>logistic</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.067</td>
<td>0.07</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>0.158</td>
<td>0.201</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>0.034</td>
<td>0.008</td>
</tr>
<tr>
<td>$f^*$</td>
<td>0.07</td>
<td>0.044</td>
</tr>
<tr>
<td>$A$</td>
<td>11.11</td>
<td>10.16</td>
</tr>
</tbody>
</table>
Table III. Posterior mean for the average age at infection and $f$ obtained from the beta-binomial models.

<table>
<thead>
<tr>
<th>Rubella</th>
<th>$f$</th>
<th>$\bar{f}$</th>
<th>$\bar{A}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.007</td>
<td>9.27</td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.016</td>
<td>9.58</td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td>0.025</td>
<td>9.85</td>
<td></td>
</tr>
<tr>
<td>0.03</td>
<td>0.033</td>
<td>10.24</td>
<td></td>
</tr>
<tr>
<td>0.04</td>
<td>0.043</td>
<td>10.63</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.053</td>
<td>11.08</td>
<td></td>
</tr>
<tr>
<td>0.06</td>
<td>0.063</td>
<td>11.55</td>
<td></td>
</tr>
<tr>
<td>Mumps</td>
<td>0.00</td>
<td>0.004</td>
<td>5.29</td>
</tr>
<tr>
<td>0.01</td>
<td>0.012</td>
<td>5.58</td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td>0.021</td>
<td>5.99</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1. Seroprevalence samples of rubella (a) and mumps (b).
Figure 2. Nonparametric estimates for the prevalence and the force of infection. Solid line: isotonic regression for the prevalence and kernel smoother for the force of infection. Dashed/dotted line: local quadratic model.
Figure 3. Posterior means for the prevalence (left panels) and the force of infection (right panels). Solid line: exponential model with $\alpha_3 = 0$, long dashed line: exponential model with $\alpha_3 > 0$, dashed line: log-logistic model.
Figure 4. Posterior means for the prevalence and the force of infection (rubella). The gray area in panel b represents the 95% credible intervals for the force of infection. NP Bayesian denotes the Beta-binomial model. The parametric model is the log logistic model which, among the parametric models, has the smallest value of DIC.
Figure 5. Posterior means for the prevalence and the force of infection (mumps). The gray area in panel b represents the 95% credible intervals for the force of infection. NP Bayesian is the Beta-binomial model. The parametric model is the exponential model with $\alpha = 0$ which, among the parametric models, has the smallest value of DIC.
Figure 6. Kernel estimates for the posterior distribution of the force of infection for mumps at ages 3.5 - 6.5.
Figure 7. Posterior mean for the force of infection for several values of $f$. Left Panel: rubella, right panel: mumps. The solid line is the force of infection estimated by local quadratic model. The numbers to the right are the values of $f$ that were used to right truncate the prior distribution of $\pi(U)$. 
Figure 8. 95% credible intervals for the average age at infection. Solid lines: Beta-binomial models, dotted line: log logistic model, long dashed line: exponential model. The numbers to the left of the credible intervals are the posterior means for $f$, $1 - \pi(U)$. 