Epidemiol. Infect. (2008), **136**, 341–351. © 2007 Cambridge University Press
doi:10.1017/S0950268807008692 Printed in the United Kingdom

Estimating the impact of vaccination using age–time-dependent incidence rates of hepatitis B

N. HENS[1]*, M. AERTS[1], Z. SHKEDY[1], P. KUNG'U KIMANI[2], M. KOJOUHOROVA[3], P. VAN DAMME[4] AND PH. BEUTELS[4]

[1] *Center for Statistics, Hasselt University, Diepenbeek, Belgium*
[2] *Kenya Institute of Medical Research, Nairobi, Kenya*
[3] *National Center of Infectious and Parasitic diseases, Department of Epidemiology, Sofia, Bulgaria*
[4] *Centre for the Evaluation of Vaccination, Epidemiology and Community Medicine, University of Antwerp, Antwerp, Belgium*

(*Accepted 12 April 2007; first published online 17 May 2007*)

SUMMARY

The objective of this study was to model the age–time-dependent incidence of hepatitis B while estimating the impact of vaccination. While stochastic models/time-series have been used before to model hepatitis B cases in the absence of knowledge on the number of susceptibles, this paper proposed using a method that fits into the generalized additive model framework. Generalized additive models with penalized regression splines are used to exploit the underlying continuity of both age and time in a flexible non-parametric way. Based on a unique case notification dataset, we have shown that the implemented immunization programme in Bulgaria resulted in a significant decrease in incidence for infants in their first year of life with 82% (79–84%). Moreover, we have shown that conditional on an assumed baseline susceptibility percentage, a smooth force-of-infection profile can be obtained from which two local maxima were observed at ages 9 and 24 years.

INTRODUCTION

Hepatitis B (HB) is a major health problem in most parts of the world. It is a DNA virus of the Hepadnaviridae family of viruses and it replicates within infected liver cells. Most of the hepatitis B virus (HBV) disease burden is due to long-term chronic sequelae of HB, which can culminate in severe inflammation of the liver, leading to cirrhosis and hepatocellular carcinoma.

Essentially a relatively virulent pathogen borne by bodily fluids such as blood, semen, vaginal fluid and in some circumstances saliva, HBV transmission can occur via multiple routes. Perinatal transmission may occur from an infected mother to her child. Horizontal transmission from person-to-person (mostly from child-to-child) may occur at any time when very small amounts of saliva or blood from an infectious person are transferred via small skin wounds such as impetigo, scabies lesions, abrasions or leg ulcers. Transmission may occur during sexual intercourse for which the rate of sexual partner change and receptive anal intercourse are important risk factors. Finally, parenteral transmission occurs when the virus is spread by penetration of the skin with an infected object, i.e. by needle stick, mucous membrane splash, tattooing, ear piercing, etc. Health-care workers and injecting drug users are generally

* Author for correspondence: Dr N. Hens, Center for Statistics, Hasselt University, Campus Diepenbeek, Agoralaan 1, 3590 Diepenbeek, Belgium.
(Email: niel.hens@uhasselt.be)

considered key risk groups for this transmission route.

As chronic HB does not become symptomatic until many years (often decades) after the infection, the link with the initial cause, infection with HBV, is often not made. The course of the infection is highly age-dependent. A symptomatic HB case is seldom seen in infected neonates or infants (<10–15%), whereas 30–35% of adults will develop an acute hepatitis subsequent to HBV infection. Thus, on the whole HB is more likely to be an asymptomatic infection [1, 2]. Carriage and chronicity of pathology (cirrhosis and liver cancer) is also age-related with more than 35–50% of neonates, infants or children developing chronic hepatitis after exposure to HB, *vs.* 6–10% in infected adults [3].

Initially it was mainly low endemic countries that introduced universal HB vaccination, because financial resources are lacking in HBV high endemic countries. Since the beginning of 2005, 168 countries have initiated universal immunization programmes in neonates, infants or adolescents [4].

In Europe, Bulgaria was one of the first countries to introduce mandatory universal immunization for all newborns, health-care workers and patients at high risk and has implemented further measures in the programme to eliminate HBV by 2020. Due to this programme a significant decrease in the disease incidence is registered, especially in the 0, 1–3 and 4–7 years age groups. The pre-universal vaccination epidemiological conditions in Bulgaria were: (1) a prevalence of HBsAg carriers of 3–5% and a seroprevalence of $\geqslant 20\%$ for HBV markers, (2) perinatal transmission of HBV infection, with 0·87% risk of creating new carriers (18·8–23·4% of HBsAg-positive pregnant women were HBeAg positive), (3) significant acute HBV infection incidence rate with up to 25 deaths per year, as well as causing chronic infections, cirrhosis and primary liver carcinoma.

In 1992, the HBV vaccine was included in the National Immunization Calendar as part of routine infant immunizations. These immunizations are mandatory and free of charge in Bulgaria (funded by the Ministry of Health) and the programme is supervised and monitored by the Ministry. During the 8-year period of universal infant immunization (started in August 1991 in Bulgaria), a total of 541 943 newborns have completed their HBV vaccination schedule (average vaccination coverage of 92·65% for the period). This vaccination coverage is comparable with the coverage of other routine infant immunizations. In addition to routine infant immunization, HBV immunization of risk groups is carried out in Bulgaria for health-care workers and medical students, as well as haemodialysis patients, haemophiliacs and HIV-positive persons.

While stochastic models/time-series have been used before to model HB counts in the absence of knowledge on the number of susceptibles [5, 6], this paper proposes to use a method that fits into the generalized additive model (GAM) framework. GAM models with penalized regression splines [7–9] are used to model age–time-dependent incidence rates of HB in Bulgaria, where underreporting was not an issue because of the rigid mandatory surveillance system. The use of GAMs facilitates multi-dimensional flexible semi-parametric modelling exploiting the natural ordering in age and time. GAM modelling for age–time-dependent incidence rates has been addressed previously to analyse cancer rates and mortality rates [10, 11]. We will apply the technique to our data and estimate the impact of vaccination on the population.

We first introduce the Bulgarian HB data and advocate the use of GAMs to model the age–time dependence in a continuous way rather than categorical, while the effect of vaccination is explicitly taken into account. Since only symptomatic cases were recorded, a correction towards asymptomatic cases is needed. Although, we lack information on the number of susceptibles, we show that conditional on an assumed baseline susceptibility percentage, incidence rates can be used to get a smooth estimated profile of the force of infection (FOI). A sensitivity analysis on this baseline susceptibility percentage showed its impact on the estimated profile.

DATA

The dataset consists of age-specific acute HB notifications, registered in Bulgaria from 1983 to 2000, while taking note of the implementation of a selective and universal infant immunization programme. At the start of the study period, the total population of Bulgaria was 8 950 144 while by 2000 it had decreased to 8 149 468. The main reasons for this reduction were an increase in emigration after 1989 and a decrease in the birth rate. The number of live births, gradually reducing after 1980, reached a minimum of 7·7/1000 population in 1997. As a result of the downward trend in the birth rate, the natural population growth (i.e. number of live birth minus the number of deaths) is

https://doi.org/10.1017/S0950268807008692 Published online by Cambridge University Press

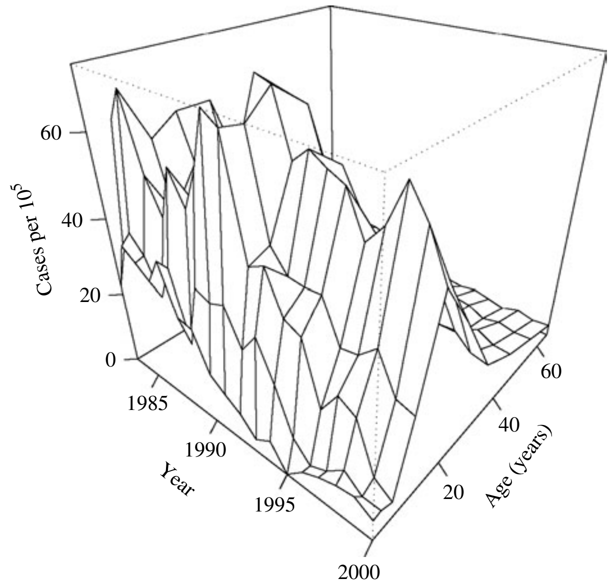

Fig. 1. Age–time perspective plot of the observed symptomatic hepatitis B rates per 10^5.

a negative value. In addition, there were changes in the age structure of the population, with an increase of the relative proportion of people aged >60 years from 16·84% in 1983 up to 21·77% in 2000 and, conversely, a reduction of the proportion of children aged 0–7 years from 11·49% in 1983 down to 7·05% in 2000.

Age–time dependence

HB has been a notifiable disease in Bulgaria since 1982. All clinically manifested acute cases with jaundice are subject to mandatory hospitalization in an infectious disease unit, following laboratory confirmation and mandatory notification and registration. The National HB Surveillance System established in 1982 requires that notification of cases of acute HB is done by age group, i.e.: 0, 1–3, 4–7, 8–13, 14–19, 20–29, 30–39, 40–49, 50–59 and ⩾60 years (www.Eurohep.net). The whole study period is divided into three parts: before the introduction of HBV immunization (1983–1987), the period of selective immunization of newborns to HBsAg-positive mothers (1988–1991), and the period of universal infant immunization (1992–2000). Figure 1 displays the acute HB rates as a function of time and age and shows a clear dependence on both.

Epidemiological and serological investigations show that various modes of transmission of HBV infection (sexual, perinatal and horizontal) have changed in importance over time. The significance of the sexual mode increased proportionally with the number of cohorts immunized against HB, to become the main mode since 1983.

Following a similar pattern over the years, the rates increased reaching a peak in the 4–7, 8–13, 14–19 or 20–29 years age groups, with most peaks observed at either 14–19 or 20–29 years. More than 50% of acute cases were in persons aged 14–29 years. Similarly, within age groups, the rates increased over time to a peak (although not monotonically) and then decreased again over the study period (Fig. 2, upper panels). For persons aged >40 years, only a slight increase of rates occurred in 1984 (40–49 and ⩾60 years) or 1985 (50–59 years), rates then gradually decreased. Children, aged between 0 and 3 years, showed the highest reduction in rate of acute cases at the end of the study period compared to the beginning, with the highest decreases starting around 1992, i.e. the time the vaccination programme started. For the 14–19 and 20–29 years age groups local peaks in the number of acute cases are observed at the beginning of the 1990s and in 1998; they are the only age groups observed to have high peaks after 1990.

Immunization programme

The vaccination programme was conducted in two stages: (1) between 1988 and 1991 immunization of newborns born to HBsAg (HB surface antigen)-positive mothers and (2) from August 1991, the Bulgarian Ministry of Health decided to administer HBV vaccine to all newborns in order to achieve a higher effectiveness. In 1992, the HBV vaccine was included in the National Immunization Calendar as part of routine infant immunizations. These immunizations are mandatory and free of charge.

All newborns are immunized according to the 0–1–6 month schedule with the first dose given during the first 24 h after birth, since pregnant women are not tested for HBsAg, without co-administration of HB immune globulin (HBIG). Coverage information about HBV vaccination is presented in Table 1. During the 8-year period of universal infant immunization, 1993–2000, a total of 541 943 newborns completed the HBV vaccination schedule. Vaccination coverage ranged from 93·54% to 97·28%, except for 1997 (due to a vaccine shortage). This vaccination coverage is comparable with the 1997

https://doi.org/10.1017/S0950268807008692 Published online by Cambridge University Press

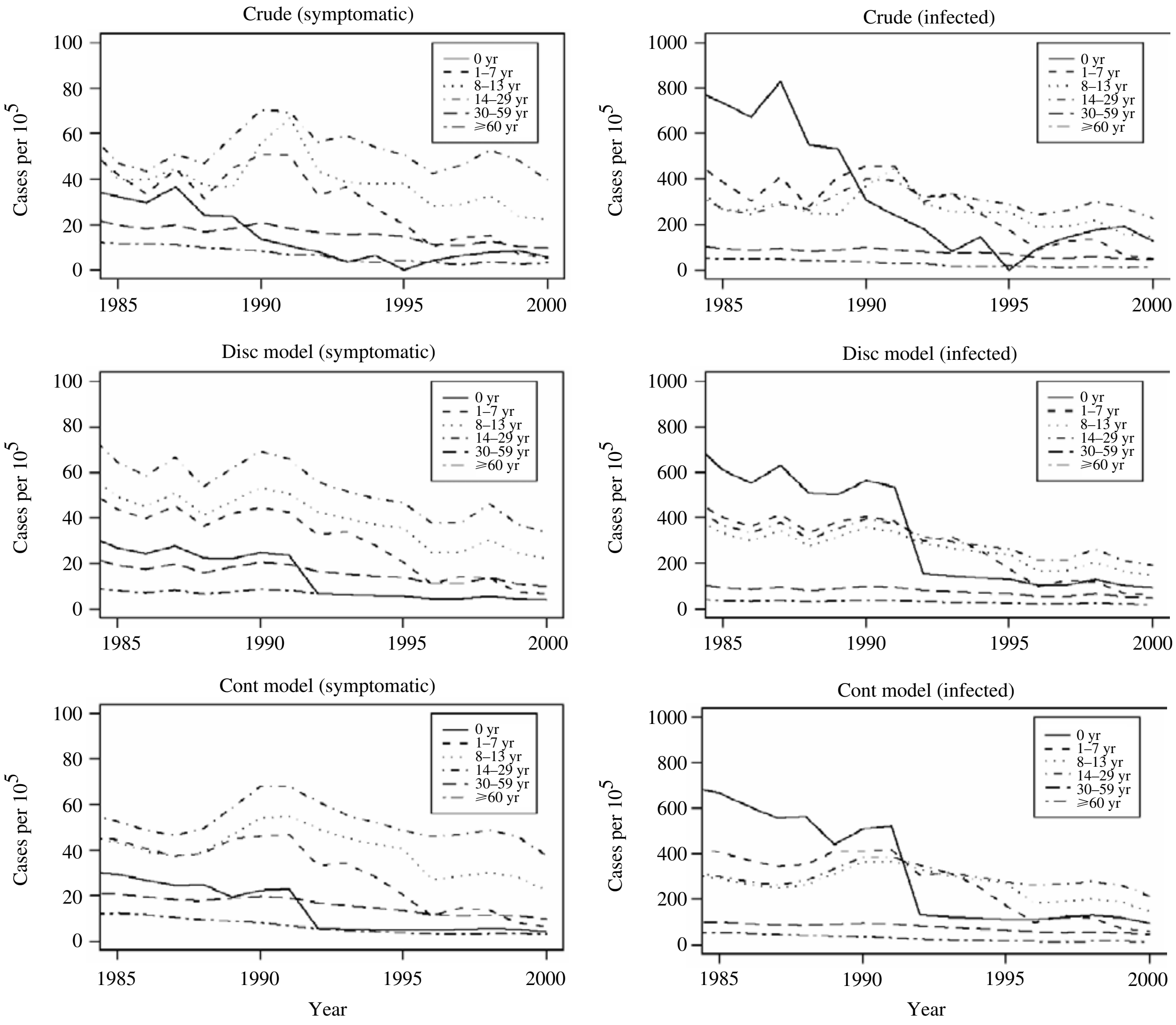

Fig. 2. Time trends for the different age categories based on the crude rates (top row), based on estimated rates using model (6) (middle row) and model (7) (lower row). Symptomatic cases are in the left column, and infected cases in the right column.

coverage of other routine infant immunizations in Bulgaria.

Asymptomatic cases

The system of registration of HBV infections does not include cases with silent or asymptomatic acute infection (without jaundice, see Introduction). The proportion of clinically manifested cases to all infections depends on the age at infection and ranges from <10% to 35%. In the forthcoming analyses, the number of symptomatic infections is the response used and the age-specific ratios are then used to derive the total number of HBV infections.

METHODS

Modelling incidence rates and the impact of vaccination

About two decades ago, the use of Poisson regression had already proved useful in modelling incidence rates. It has the attractive property of allowing postulated aetiological mechanisms of exposures and/or disease expression characteristics to be linked to the observed rates [12].

Poisson regression models are part of the generalized linear model (GLM) framework of McCullagh & Nelder [13], where the response is a Poisson random variable of which the mean is related to a systematic

Table 1. *Hepatitis B vaccination coverage in infants in Bulgaria after introduction of universal immunization (1993–2000)*

Year	1993	1994	1995	1996	1997	1998	1999	2000
HBV3 coverage (%)	95·67	94·19	95·45	93·54	77·18	97·06	97·28	93·67

HBV3 coverage denotes the coverage for a completed three-dose hepatitis B vaccination schedule.

Table 2. *Universal (U) and selective (S) immunization programmes with dummies indicating the different proportion immunized*

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
0	S_0	S_0	S_0	U_0	U_0	U_0	U_0	U_0	U_0	U_0	U_0	U_0
1–3		S_1	S_2	S_0	U_1	U_2	U_0	U_0	U_0	U_0	U_0	U_0
4–7				S_3	S_4	S_5	S_0	U_3	U_4	U_5	U_0	U_0
8–13								S_6	S_7	S_8	S_9	U_6

component via a link function. The systematic component for a GLM specifies the explanatory variables used in a linear predictor function.

The data are number of acute cases of HB for persons in a specified age group for a given year. Let us denote Y_{ij}, the number of cases in age group i ($i=1$, …, 10) at year j ($j=1$, …, 18), referred to hereafter as bins, being Poisson with mean $E(Y_{ij})=\theta_{ij}$, where

$$E(Y_{ij})=N_{ij}\exp(\mu_{ij}). \qquad (1)$$

Here N_{ij} is the number of persons at risk (population) in the ijth bin. N_{ij} is included in the mean structure to account for the different population sizes in the bins (demographic changes). We consider an additive form for the linear predictor term μ_{ij} of the following form

$$\mu_{ij}=\mu_0+\alpha_i+\beta_j+\sum_{k=0}^{6}\gamma_k I_{ij}^{U_k}+\sum_{l=0}^{9}\delta_l I_{ij}^{S_l}, \qquad (2)$$

where α_i represents the effect of age group i, β_j the effect of year j. $I_{ij}^{U_k}$ represents an indicator variable taking value 1 if the kth universal immunization programme U_k ($k=0$, …, 6) took place in age group i and year j and 0 otherwise in accordance with Table 2. Similarly, $I_{ij}^{S_l}$ is an indicator variable for the lth selective immunization programme, $l=0$, …, 9. When the corresponding coefficients $\gamma_k=\gamma$, for all k ($\delta_l=\delta$ for all l), there is no distinction between different proportions immunized for the universal (selective) immunization programme.

Using model (2), age and year are treated as categorical variables, ignoring the underlying natural ordering of these variables. An alternative is to include year and age as continuous variables in the model. A first approach is to supplement model (2) with an interaction term $\gamma_i v_j$, $\gamma_j u_i$ or $\gamma u_i v_j$, where u_i denotes the midpoint of the ith age category [14] and v_j denotes year j. Adding $\gamma u_i v_j$, we assume the interaction to have a linear effect on the incidence rates, while $\gamma_i v_j$ and $\gamma_j u_i$ allow for more flexibility in the interaction at the cost of the number of parameters used. Note that we cannot add a discrete interaction term γ_{ij} to model (2) since this would lead to an overparametrized model. A second approach to exploit the underlying continuity is to treat age (midpoints) and year as continuous predictors. Specifying the parametric functional relationship, including interactions, to relate these predictors to $E(Y_{ij})$ is, however, difficult.

A generalization, replacing the linear predictor of a GLM by smooth functions of the predictors, splines, is provided by using GAMs as originally introduced by Hastie & Tibshirani [15]. Splines are generally defined as piecewise polynomials in which curve (or line) segments are constructed individually and pieced together at what are called 'knots'. A large number of knots allows more flexible forms to be taken but results in a non-smooth function. Using penalized regression splines, a penalty on the roughness of the corresponding coefficient vector is set. This penalty, controlled by a smoothing parameter regulates the trade-off between the fit of the data and the smoothness. There are many methods to select the optimal smoothing parameter, e.g. the Akaike Information Criterion (AIC [16]) and generalized cross-validation (GCV [17]).

The GAM methodology was further developed by Wood (see e.g. [8, 18, 19]), who has done a great deal

https://doi.org/10.1017/S0950268807008692 Published online by Cambridge University Press

of work on the application of the technique using penalized regression splines [7, 8, 20, 21]. The motivation of Wood's work was to overcome the difficulties associated with model selection and inference when backfitting with linear smoothers [22]. The mathematically elegant work of Wahba [21] on generalized spline smoothing provides a rigorous framework for model selection and inference with GAMs. A 'middle way' between these approaches was the use of penalized regression splines to construct GAMs. The availability of the R package *mgcv* [23, 24] has made the use of GAMs very popular. The systematic component of the GAM version of model (2) is given by

$$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_i) + te(a_i, y_i) + \sum_{k=0}^{6} \gamma_k I_{ij}^{U_k} + \sum_{l=0}^{9} \delta_l I_{ij}^{S_l}, \quad (3)$$

where f and g denote penalized regression splines of predictor age (a) and year (y), respectively and $te(a, y)$ is a tensor-product spline, which can be looked upon as a smooth interaction between two variables [21]. The GAM method uses GCV to select the smoothing parameter.

Note that in model (3), we used the usual thin plate spline penalty as the measure of smoothness for $f(a)$ and $g(y)$ and from a model-building perspective added to this a cubic tensor product spline, i.e. the proposed model uses different measures of smoothness, and is therefore different from the model with merely $te(a, y)$.

A key feature of the Poisson distribution is that its variance equals its mean. In practice, count observations often exhibit variability exceeding that predicted by Poisson. This phenomenon is called overdispersion and is often caused by subject heterogeneity. There are several ways to deal with overdispersion [25], e.g. the use of a scaling factor and random-effects models. The approach presented here replaces the Poisson distribution by the negative binomial which is a gamma mixture of Poisson distributions. A negative binomial distribution has mean θ and variance $\theta + \theta^2/k$ where $1/k$ is often referred to as the dispersion parameter. As $1/k \to 0$, the negative binomial distribution converges to the Poisson distribution with mean and variance θ.

Selecting the optimal model among the set of submodels from models (2) and (3), is done using the AIC criterion [15, 16]. The AIC value of a model is given by $-2\text{LL} + 2K$, where LL denotes the log-likelihood and K the number of (effective) parameters in that model. The model with the lowest AIC value is chosen to be the optimal model among the set of models under consideration, i.e. the model with optimal balance between goodness of fit (measured by -2LL) and complexity (measured by $2K$).

Deriving a FOI profile

A fundamental parameter, describing infectious disease dynamics is the FOI, i.e. the rate at which a susceptible individual becomes infected. This rate is known to be age-dependent and different methods have been developed to estimate it from serological data. It is not possible to estimate the FOI from case-notification data alone, due to the lack of knowledge on susceptibility in the population at hand. However, starting from an assumed percentage of susceptibility for newborns, it is possible to obtain a conditional FOI estimate. Varying the percentage of susceptibility then results in a sensitivity analysis on the estimated curve.

As is done for serological data, we assume time homogeneity, i.e. we assume a cohort passes through different age classes ignoring the effect of changes within age classes over time. One could state this to be too strong an assumption. On the one hand, we only have time-dependent data for the period 1983–2000, which is too limited for HB, since the disease is not only transmitted horizontally, typically around the age of 10 years, but also sexually, typically around ages 20–30 years. On the other hand, the estimates of FOI based on the year-specific data over a period of 18 years, can give us a good idea of the variability on the estimated curve over time.

Fixing time t, the aim is to calculate the FOI $\lambda(a)$, given the incidence rates $I_a = X_a/N_a$, where X_a is the number of infections and N_a is the population size in age class a at time t (omitted from notation). The FOI is given by

$$\lambda(a) = X_a/S_a, \quad (4)$$

where S_a ($a = 1, 2, \ldots$) is the number of susceptibles at age a. S_a can be calculated recursively as

$$S_a = \frac{S_{a-1} - X_{a-1}}{N_{a-1}} N_a, \quad (5)$$

with $S_0 = p_s N_0 \equiv p_s N_1$ and $X_0 \equiv 0$. p_s denotes the proportion susceptible at birth, referred to as the baseline susceptibility proportion hereafter. Note that equation (5) takes the age distribution of the population into account.

https://doi.org/10.1017/S0950268807008692 Published online by Cambridge University Press

Table 3. *Candidate models together with their empirical degrees of freedom (edf) and AIC value based on model (3)*

Model	edf	AIC
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0$	178·0	22761·93
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i)$	170·2	2019·95
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + g(y_i)$	176·2	2198·80
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_i)$	165·0	1883·16
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_j) + te(a_i, y_j)$	147·3	1713·12
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_j) + te(a_i, y_j) + \gamma I_{ij}^{U}$	146·5	1714·47
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_i) + te(a_i, y_i) + \sum_{k=0}^{6} \gamma_k I_{ij}^{U_k}$	143·2	1659·76
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_i) + te(a_i, y_i) + \delta I_{ij}^{S}$	146·1	1714·48
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_i) + te(a_i, y_i) + \sum_{l=0}^{9} \delta_l I_{ij}^{S_l}$	138·0	1730·36
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_i) + te(a_i, y_i) + \gamma I_{ij}^{U} + \delta I_{ij}^{S}$	144·8	1710·27
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_i) + te(a_i, y_i) + \gamma I_{ij}^{U} + \sum_{l=0}^{9} \delta_l I_{ij}^{S_l}$	136·6	1725·54
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_i) + te(a_i, y_j) + \sum_{k=0}^{6} \gamma_k I_{ij}^{U_k} + \delta I_{ij}^{S}$	141·4	1633·92
$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_i) + te(a_i, y_j) + \sum_{k=0}^{6} \gamma_k I_{ij}^{U_k} + \sum_{l=0}^{9} \delta_l I_{ij}^{S_l}$	132·0	1637·67

Similar to modelling the incidence rates, using model (3), we can model the susceptibility rates calculated from the crude data using equation (5) conditional on the assumed proportion of susceptibility p_S. Using the estimated incidence and susceptibility rates the FOI is given by equation (4). p_S is unknown and could optimally be estimated from serological data (see e.g. [26]). Since these data are not available, we let p_s vary over a range of values and look at the effect on the FOI estimate.

To eliminate the influence of vaccination from our FOI estimate, the number of susceptibles and infections is estimated while putting the immunization effects in the estimated models to zero. In this way, one mimics the situation where no immunization would have occurred.

RESULTS

Modelling incidence rates and the impact of vaccination in Bulgaria

From all candidate models starting from model (2), the model with a minimal AIC value of 1789·83 (on 144 D.F.)

$$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + \alpha_i + \beta_j + \sum_{k=0}^{6} \gamma_k I_{ij}^{U_k} + \delta I_{ij}^{S}, \tag{6}$$

includes both the bin-specific universal and the selective immunization programme.

Exploiting the underlying continuity by including $\gamma_i v_j$, $\gamma_j u_i$ or $\gamma u_i v_j$ into model (6), using the midpoints (0·5, 2·5, 6, 11, 17, 25, 35, 45, 55 and 65 years) for the age categories (see e.g. [14]), improves the AIC value to 1652·64, 1653·26 and 1678·33, respectively. Further exploiting the underlying natural ordering of age and time, candidate models based on a GAM model, as described by model (3), were used. In Table 3, candidate models are shown again with their (empirical) degrees of freedom (edf) and AIC value. Here, *f*() and *g*() denote penalized regression splines and *te*(,) a tensor-product spline (the smooth version of an interaction).

The model with lowest AIC value is the continuous version of model (6), where the bin-specific universal immunization programme and the selective immunization programme are included again

$$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_j) + te(a_i, y_i) + \sum_{k=0}^{6} \gamma_k I_{ij}^{U_k} + \delta I_{ij}^{S}. \tag{7}$$

The AIC value of the latter model is substantially lower than that of model (6). Trying to reduce the model leaving out the tensor-product spline worsens the fit, so does leaving out the one-dimensional spline functions of age and year. The overdispersion parameter k in model (7) was estimated at *50* (P value 0·0002). In Figure 2 (left panel), time trends for the different age categories are shown for both models together with the observed rates. Using model (7) produces smoother curves compared to using model (6). The sudden drop in 1992 for the youngest age class (0 years) deviates from the corresponding crude profile. This effect is enlarged when applying the age-specific factor to obtain an estimate of the total

https://doi.org/10.1017/S0950268807008692 Published online by Cambridge University Press

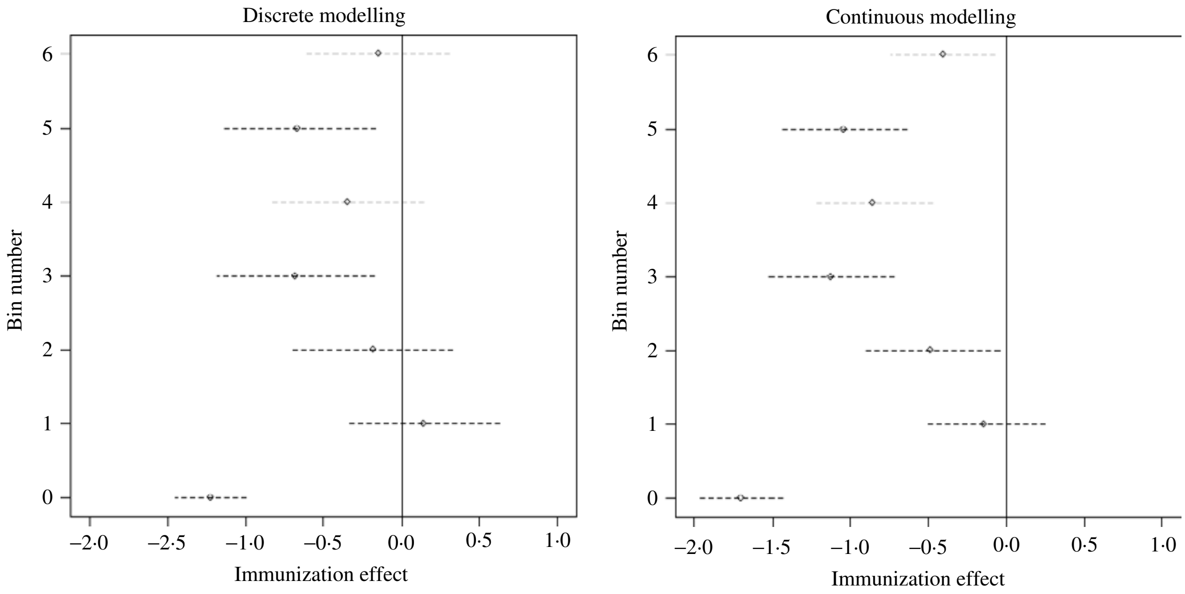

Fig. 3. Parameter estimates and confidence intervals for the bin-specific universal immunization programme according to model (6) (left panel) and model (7) (right panel).

number of infections and is a direct result of the model approach where $I_{ij}^{U_0}$ in model (7) is responsible for this behaviour since it comprises not only the implemented universal immunization programme in the 0 years age group from 1992 onwards but also in age groups 1–3 years and 4–7 years from, respectively, 1995 and 1999 onwards (Table 2). Alternatively, using the model

$$\log(\theta_{ij}) = \log(N_{ij}) + \mu_0 + f(a_i) + g(y_i) + te(a_i, y_i) + \gamma_{01} I_{ij}^{U_{01}} + \gamma_{02} I_{ij}^{U_{02}} + \gamma_{03} I_{ij}^{U_{03}} + \sum_{k=1}^{6} \gamma_k I_{ij}^{U_k} + \delta I_{ij}^{S},$$

where $\gamma_{01} I_{ij}^{U_{01}} + \gamma_{02} I_{ij}^{U_{02}} + \gamma_{03} I_{ij}^{U_{03}}$ replaces $\gamma_0 I_{ij}^{U_0}$ in model (7) to distinguish between these age categories resulted in a fitted profile close to the crude one for the 0 years age group but has a corresponding AIC value of 1644·57 which is considerably higher than the AIC value 1633·92 of model (7). Model (7) is therefore retained as the best model among this set of candidate models.

In Figure 3 bin-specific parameter estimates and 95% confidence intervals (CIs) are shown for the universal programme for both model (6) (left panel) and model (7) (right panel). While the circle represents the point estimate, the dashed line represents the 95% CI. Using the underlying continuous nature of both age and time results in a more dramatic immunization effect. The implementation of the mandatory vaccination resulted in a significant decrease in incidence among infants in their first year of life with 82% [based on model (7) with 95% CI 79–84%], showing the benefits of immunization. One has to take care in using these estimates to make predictions for the situation where no immunization had occurred. Indeed, since we do not model the underlying dynamic transmission process, and our observations are based on acute symptomatic infections only, our vaccine-free estimates do not account for all possible aspects of herd immunity. Herd immunity reduces the FOI in non-vaccinated individuals (i.e. reduces their risk of infection per unit time), and therefore increases the average age at infection of residual infections [27, 28]. For HB, this age shift will lead to proportionately more acute infections, but fewer chronic infections. Our model is based on observations related to acute infections only, which we inflated to obtain the total number of infections, by applying an independent age-specific factor to these observations. Therefore, our model underestimates the infection-reducing impact of vaccination. However, since our model combines age- and time-specific observations on both incidence and vaccination status, we expect this implicit age-related distortion caused by herd immunity to be limited.

Turning to the infected cases, i.e. correcting for the asymptomatic cases by multiplying with age-specific ratios, model (7) again produces smoother curves

https://doi.org/10.1017/S0950268807008692 Published online by Cambridge University Press

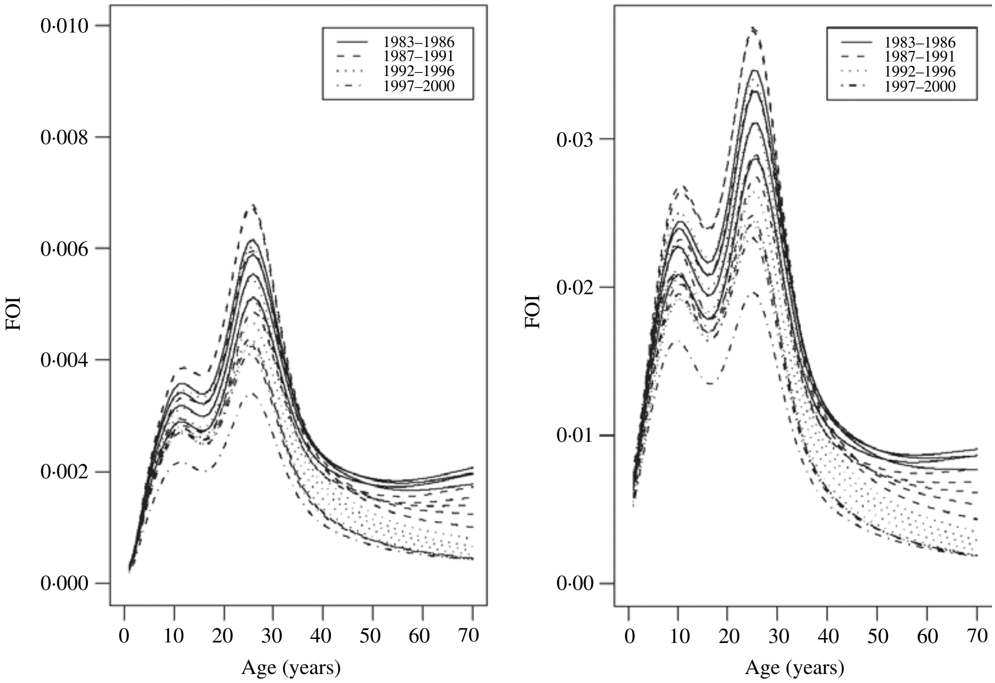

Fig. 4. Force-of-infection (FOI) profiles for the whole period 1983–2000 based on the estimated symptomatic hepatitis B (HB) cases (left panel) and estimated HB infections (right panel).

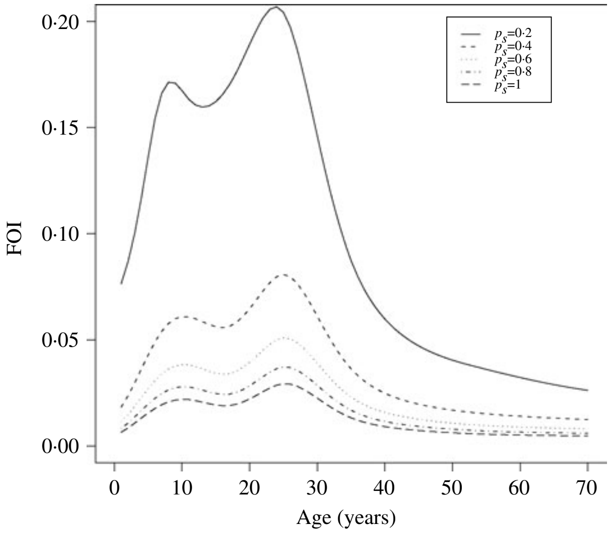

Fig. 5. Plot of the aggregated force-of-infection (FOI) profiles for different baseline susceptibility percentages.

(Figure 2, right panel). The correction for asymptomatic cases clearly shows the benefit of immunization. The number of infected children aged 0–1 year decreases enormously in 1992, the year universal immunization programmes started. Analogously, decreases through the years are noticed for the successive age categories showing also a herd immunity effect, especially for the lower age categories.

Deriving a FOI profile in Bulgaria

Putting immunization effects to zero to mimic the effect of no vaccination, the resulting incidence and susceptibility rates are used to derive the FOI using equations (4) and (5) for different baseline proportions of susceptibility p_s.

In Figure 4 the estimated year-specific FOI profiles (conditional on $p_s = 1$) are shown for acute cases (left panel) and infected cases (right panel). Figure 5 shows aggregated FOI curves based on different values of p_s, the baseline susceptibility percentage. While p_s ranges from 0·2 to 1·0, the shape of the FOI curve remains about the same, while the magnitude decreases with increasing percentage susceptibility.

The age-specific FOI profiles give two local maxima which are located around the age of 9 years and 24 years, respectively. These findings have also been described in other countries [29]. The first local maximum illustrates the importance of horizontal transmission between children. The second local maximum shows the importance of sexual transmission and potentially also parenteral transmission by, e.g. needle sharing for injecting drug use among young adults. The importance of horizontal transmission should not be underestimated, and becomes particularly apparent when observing the true FOI (i.e. based on the total number of infections).

https://doi.org/10.1017/S0950268807008692 Published online by Cambridge University Press

DISCUSSION

In this paper, the evolution of HB in Bulgaria is described for the period 1983–2000. Bulgaria was one of the first countries in Europe to introduce mandatory universal immunization for all newborns and data recorded included age-category-specific acute symptomatic cases and population sizes. This setting is unique, since accurate information of the number of acute clinically manifested cases is obtained, making underreporting no longer an issue. From there, by using approaches new to the field of infectious diseases, not only incidence rates were estimated, but also qualitative insights in the FOI were obtained. A maximal FOI was found around ages 9 years and 24 years, illustrating the importance of horizontal transmission between children and of sexual and parenteral (e.g. drug-related) transmission among young adults.

A GAM with penalized splines on the number of acute cases was used to smooth over the different age categories, exploiting the underlying continuous nature of age and time. Age-specific ratios were used to include asymptomatic cases, resulting in a four- to ten-fold increase in incidence. The proposed method supplements the existing stochastic models/time-series that have been proposed to analyse case notification data and is, thanks to existing open source software, easy to apply.

This research illustrates that the HB vaccination programme in Bulgaria has had a rapid and substantial impact on HBV incidence and thus can be considered as successful. Moreover, the approach presented in this paper, shows that case notification data can be used to obtain qualitative insights in the behaviour of the FOI with age, even in the presence of a vaccination programme.

ACKNOWLEDGEMENTS

We thank both referees and the associate editor for valuable remarks leading to an improved presentation. We gratefully acknowledge support from the Fund of Scientific Research (FWO grant G.0393.04) Flanders, Belgium and from the IAP research network no. P5/24 of the Belgian Government (Belgian Science Policy).

DECLARATION OF INTEREST

None.

REFERENCES

1. **McMahon BJ, Alward WLM, Hall DB.** Acute hepatitis B virus infection: relation of age to the clinical expression of disease and subsequent development of the carrier state. *Journal of Infectious Diseases* 1985; **15**: 604–609.
2. **Shapiro CN.** Epidemiology of hepatitis B. *Pediatric Infectious Disease Journal* 1993; **12**: 443–447.
3. **Edmunds WJ, Medley GF, Nokes DJ.** The influence of age on the development of the hepatitis B carrier state. *Proceedings of the Royal Society of London Series B, Biological Sciences* 1993; **253**: 197–201.
4. **FizSimons D, *et al.*** Strengthening immunization systems and introduction of hepatitis B vaccine in central and Eastern Europe and the Newly Independent States. Viral Hepatitis Prevention Board, University of Antwerp. 2005.
5. **Held L, *et al.*** A statistical framework for the analysis of multivariate infectious disease surveillance data. *Statistical Modelling* 2005; **5**: 187–199.
6. **Held L, *et al.*** A two component model for counts of infectious diseases. *Biostatistics* 2006; **7**: 422–437.
7. **Marx BD, Eilers PHC.** Direct generalized additive modelling with penalized likelihood. *Computational Statistics and Data Analysis* 1998; **28**: 193–209.
8. **Wood SN.** Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B* 2000; **62**: 413–428.
9. **Aerts M, *et al.*** Some theory for penalized spline additive models. *Journal of Statistical Planning and Inference* 2002; **103**: 455–470.
10. **Thurston S, *et al.*** Negative binomial additive models. *Biometrics* 2000; **56**: 139–144.
11. **Currie I, *et al.*** Smoothing and forecasting mortality rates. *Statistical Modelling* 2004; **4**: 279–298.
12. **Frome EL, Checkoway H.** Epidemiologic programs for computers and calculators. Use of Poisson regression models in estimating incidence rates and ratios. *American Journal of Epidemiology* 1985; **121**: 309–323.
13. **McCullagh P, Nelder J.** *Generalized Linear Models*. Chapman & Hall, 1989.
14. **Smith L, *et al.*** Spline interpolation for demographic variables: the monotonicity problem. *Journal of Population Research* 2004; **21**: 95–98.
15. **Hastie T, Tibshirani R.** *Generalized Additive Models*. Chapman and Hall, 1990.
16. **Akaike H.** Information theory as an extension of the maximum likelihood principle. In: Petrov B, Csaki F, eds. *Second International Symposium on Information Theory*. Budapest: Akademia Kiado, 1973, pp. 267–281.
17. **Craven P, Wahba G.** Smoothing noisy data with spline functions. *Numerische Mathematik* 1979; **31**: 377–403.
18. **Wood SN.** mgcv: GAMs and generalized ridge regression for R. *R News* 2001; **1**: 20–25.
19. **Wood SN.** Thin plate regression splines. *Journal of the Royal Statistical Society* 2003; **65**: 95–114.

https://doi.org/10.1017/S0950268807008692 Published online by Cambridge University Press

20. **Wahba G.** Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy data. In: Cheney W, ed. *Approximation Theory III*. New York: Academic Press, 1980, pp. 905–912.
21. **Wahba G.** *Spline Models for Observational Data*, CBMS-NSF series. Philadelphia: SIAM, 1990.
22. **Hastie T, Tibshirani R.** Generalized additive models: some applications. *Journal of the American Statistical Association* 1987; **82**: 371–386.
23. **RDC Team R.** A language and environment for statistical computing, 2004.
24. **Wood SN.** Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* 2004; **99**: 673–686.
25. **Agresti A.** *Categorical Data Analysis*. Wiley & Sons, 2002.
26. **Wallinga J, *et al.*** Estimation of measles reproduction ratios and prospects for elimination of measles by vaccination in some Western European countries. *Epidemiology and Infection* 2001; **127**: 281–295.
27. **Anderson RM, May RM.** Age-related changes in the rate of disease transmission: implications for the design of vaccination programmes. *Journal of Hygiene* (*London*) 1985; **94**: 365–435.
28. **Fine PEM.** Herd immunity: history, theory, practice. *Epidemiologic Reviews* 1993; **15**: 265–302.
29. **Beutels P, *et al.*** Hepatitis B in St Petersburg, Russia (1994–1999): incidence, prevalence and force of infection. *Journal of Viral Hepatitis* 2003; **10**: 141–149.

https://doi.org/10.1017/S0950268807008692 Published online by Cambridge University Press