Published in final edited form as:

Environmetrics. 2008 June 2; 20(3): 235–248. doi:10.1002/env.923.

Time-distributed effect of exposure and infectious outbreaks

Elena N. Naumova^{1,*†‡} and Ian B. MacNeill²

¹ Department of Public Health and Family Medicine, Tufts University School of Medicine, Boston, MA, USA

² University of Western Ontario, London, Canada

SUMMARY

Extreme weather affects the timing and intensity of infectious outbreaks, the resurgence and redistribution of infections, and it causes disturbances in human-environment interactions. Environmental stressors with high thermoregulatory demands require susceptible populations to undergo physiological adaptive processes potentially compromising immune function and increasing susceptibility to infection. In assessing associations between environmental exposures and infectious diseases, failure to account for a latent period between time of exposure and time of disease manifestation may lead to severe underestimation of the effects. In a population, health effects of an episode of exposure are distributed over a range of time lags. To consider such time-distributed lags is a challenging task given that the length of a latent period varies from hours to months and depends on the type of pathogen, individual susceptibility to the pathogen, dose of exposure, route of transmission, and many other factors. The two main objectives of this communication are to introduce an approach to modeling time-distributed effect of exposures to infection cases and to demonstrate this approach in an analysis of the association between high ambient temperature and daily incidence of enterically transmitted infections. The study is supplemented with extensive simulations to examine model sensitivity to response magnitude, exposure frequency, and extent of latent period.

Keywords

distributed lag model; poisson regression; time series analysis; outbreak signature; Monte Carlo simulations; waterborne and food-borne infections

1. INTRODUCTION

Environmental stressors with high thermoregulatory demands require susceptible populations to undergo physiological adaptive processes potentially compromising immune functions and increasing susceptibility to infection. Weather can directly impact disease transmission through effects on the disease agents, and indirectly via change in ecology and human behavior. High temperature leads to higher water consumption, promotes outdoor swimming, camping, recreational water use, and other outdoor activities. It also promotes the replication of pathogenic bacteria in food and on food preparation surfaces (Hobbs and Roberts, 1987; Kapperud *et al.*, 1992; Kusumaningrum *et al.*, 2003; Neimann *et al.*, 2003). Excessive heat can also exacerbate gastrointestinal infection due to heat-related dehydration,

*Correspondence to: E. N. Naumova, Department of Public Health and Family Medicine, Tufts University School of Medicine, 136 Harrison Ave, Boston, MA 02111 USA.

†elena.naumova@tufts.edu

‡Associate Professor.

as reported in the deadly European heat wave of August 2003 (Fuchs, 200; Brock, 200). A comprehensive study from the UK (Bentham and Langford, 2001) describes a short-term link between temperature and food poisoning (often caused by *Salmonella* (Sackett, 1991; Cowden *et al.*, 1995; Djuretic *et al.*, 1997; Evans *et al.*, 1998)). Climate change could also lead to changes in rates of food poisoning (Bentham and Langford, 1995). The majority of waterborne outbreaks of enteric diseases occur in the summer months (Joce *et al.*, 1991; Barwick *et al.*, 2000; Fleming *et al.*, 2000; Lee *et al.*, 2002). Our own research shows that *Salmonella* and *Campylobacter* infections in the temperate climate of Massachusetts closely followed the seasonal ambient temperature curve.

In environmental epidemiology, time-series studies for assessing associations between environmental exposures and health outcomes have gained popularity over the last decade. Large databases of billing records for health care service utilization and increased implementation of monitoring systems provide long time series of events and measurements at fine temporal resolution. Studies of air quality and health have a history of using models that are able to take into account time-distributed effects of exposure (Wyzga, 1978; Schwartz, 2000; Huang *et al.*, 2004; Welty and Zeger, 2005). These types of models may have a greater potential for assessing associations between environmental exposures and infectious diseases because of delayed disease manifestation or incubation time.

Meteorological factors, such as extreme precipitation events (Naumova *et al.*, 2005a) or ambient temperature (Bentham and Langford, 2001), may drive seasonal patterns in the probability of exposure and transmission of waterborne pathogens. Understanding these relationships is important for forecasting disease trends, design of warning systems, developing efficient outbreak detection algorithms, and regulatory practices for preventing enterically transmitted infections. In temperate climates, seasonality in waterborne enteric infections characterized by periods of low endemic levels and clusters of outbreaks, which form a unique seasonal pattern, is a well-known phenomenon (Naumova *et al.*, in press). In Massachusetts, enteric infection caused by *Cryptosporidium* exhibits a seasonal increase in warm months that is shifted toward fall (Naumova *et al.*, 2000). Systematic temporal fluctuations in diseases that have similar sources of exposure, routes of transmissions, and clinical manifestations might imply a presence of environmental factors that synchronize seasonal curves. In assessing associations between extreme ambient temperature and infectious diseases, failure to account for a latent period between time of exposure and disease manifestation may lead to severe underestimation of the effects on the exposed population. As Chetty (1971) pointed out, one would like to formulate a statistical model that does not pose serious estimation problems and satisfactorily reflects the underlying theory, in our case—the epidemiology of an infectious process in a community.

Models that account for time-distributed delays are widely implemented in many practical econometric applications. The general form of a linear distributed lag model is

$$Y_t = a + \sum_{i=0}^{\infty} \beta_i X_{t-i} + \varepsilon_t$$

where any change in X_t will have an effect on $E[Y_t]$ in all later periods. The term β_i is the i th coefficient (or reaction coefficient in econometric literature), and it is usually assumed that

$\lim_{i \rightarrow \infty} \beta_i = 0$ and $\sum_{i=0}^{\infty} \beta_i = \beta < \infty$. If we also assume that after some period of time, K , the effect of X_t practically vanishes, then the model reduces to a finite distributed lag model for which the upper limit of the summation is K . The upper limit of summation defines the

number of parameters in the model. In order to be parsimonious, the parameters in a distributed lag model have to be functionally related (Jorgensen, 1966; Schwartz, 2000), and form a relationship defined by different types of time-distributed lag structures. A geometric lag model that has a lag structure with an exponential decay was introduced by Koyck (1954) and assumes that $\beta_i = \alpha \lambda^i$, where $0 < \lambda < 1$. Solow (1960) suggested a flexible weighting scheme, where β_i are generated by a Pascal distribution defined as

$$\beta_i = \alpha \binom{r+i-1}{i} (1-\lambda)^r \beta_{i-1}, 0 < \lambda < 1, \forall i, r > 0 \quad (1)$$

if $\alpha = 1$, then

$$\beta_i = \frac{r+i-1}{i} \lambda^r \beta_{i-1} \quad (2)$$

and if $r = 1$, then the Solow's weighting scheme reduced to the Koyck's model. Another model with a flexible lag structure was proposed by Almon (1965), which is also known as a model with "interpolation distribution" for parameters. In this model, the coefficients β_i are approximated by polynomials in the lag i of degree $p < m$, with $p + 1$ parameters, so that

$$\beta_i = \sum_{k=0}^p \alpha_k t^k. \quad (3)$$

this polynomial approximation provides a variety of shapes for a lag structure.

Intuitively, this technique can be used to describe temporal patterns in manifestations of infectious diseases. In our studies of the Milwaukee outbreak of cryptosporidiosis we demonstrated a delayed effect of drinking water turbidity on the incidence of non-specific gastroenteritis using models with autoregressive components for endogenous and exogenous variables (Morris *et al.*, 1998; Naumova *et al.*, 2003). In those analyses we assumed that a uniform approximation to the lag structure was sufficient for the existing outcome and developed an approach to visualize the lagged relations using the temporal exposure response surfaces (TERS), which allowed us to describe the lagged structure of a health outcome in response to elevated water quality parameters (Chetty, 1971). The shape of lagged responses to the spikes in turbidity resembled the shape of a distribution that would be characteristic for incubation times of cryptosporidiosis, most prominently at 6–9 days after the peak in exposure. Further analysis of reported distributions of incubation periods for cryptosporidiosis supported the idea of considering the knowledge of distributions for incubation periods in the time series analysis (Naumova *et al.*, 1999). Recently, we developed a distributed lag model in which the weights for each lag were estimated from the Poisson distribution with a parameter corresponding to a disease-specific mean incubation period in the population (Egorov *et al.*, 2003). This model was applied to assess relations between daily variations in effluent water turbidity and enteric illness. The potential for the proposed approach was further evident in our development of a system of adaptive algorithms for detecting and forecasting waterborne outbreaks (Naumova and MacNeill, 2005; Naumova *et al.*, 2005b), where we introduced the notion of a disease "signature" to

express manifestation of infection in response to an agent and incorporated this disease “signature” into short-term and long-term forecasts.

The two main objectives of this communication are to introduce an approach for modeling of the time-distributed effect of exposures to infection cases and to demonstrate this approach in the analysis of the association between high ambient temperature and daily incidence of enterically transmitted infections. As an example, we examined the effect of episodes of ambient temperature over 90°F on reported waterborne cryptosporidiosis in Massachusetts. The study has been supplemented with extensive simulations to examine model sensitivity to response magnitude, exposure frequency, and extent of latent period.

2. MODELING A SINGLE-SOURCE OUTBREAK

To outline a rationale for statistical modeling of a potentially waterborne infection, we make three major assumptions based on disease epidemiology (Koyck, 1954). We assume that an infection is always present in a community at a low level. Following an episode of elevated environmental exposure to a causal agent, the incidence of a food- or water-borne infection may increase substantially. This increase in the disease rate above an endemic level is localized in time and space, meaning that an outbreak has an onset, magnitude, and duration consistent with the distribution of incubation periods of this pathogen in the population. In this study we limit our framework to a single-source outbreak with inconsequential secondary amplification.

We assume that the occurrence of waterborne infections in a community can be described as a combination of two stochastic processes, endemic, and epidemic. The numbers of events or counts observed in regularly spaced intervals over some period T are the realizations of these processes. An endemic process, Y_e , is observable only in a time period free of outbreak and is assumed to be stationary for a given community. The process associated with an outbreak is denoted Y_o . During an outbreak period the number of events on day t starts to increase after time d_0 since an episode of exposure, peaks at the time corresponding to the modal value for the distribution of incubation periods for the responsible pathogen, and then decays by day d_k to a pre-specified level. Let $D_t = 0$, if $0 < t < d_0$ or $t > d_k$ and $D_t = 1$, if $d_0 \leq t \leq d_k$. Then the observed number of events on any given day t is $Y_t = Y_e + D_t Y_o$.

We assume that the number of endemic cases on day t in a time period free of outbreak, Y_e , follows a Poisson distribution with parameter, λ , reflecting an average endemic level, and that Y_e is a sequence of i.i.d random variables. Let Y_{ot} be the number of cases on day t , attributable to the outbreak, so $d_0 \leq t \leq d_k$. Let us assume that an additional n people fall ill during the outbreak time period. Let Z_i be the incubation time (time from day of exposure to day of onset of symptoms) for person i . Under the assumption that the incubation time Z_i follows a Poisson distribution, the probability of falling ill on day $t = d_0 + z$ for those who were exposed on day d_0 is $P(Z_i = z) = e^{-\nu} \nu^z / z!$, where $z = 0, 1, 2, \dots$, and ν is the population mean incubation period for the responsible pathogen. Then, the number of cases on day t that are due to the outbreak will depend upon n , Bernoulli sampling, and the probability of becoming ill on day t . Thus, the number of cases on day t is $Y_{ot} = \sum_{i=1}^n I(Z_i = t - d_0)$ where I is an indicator function and $I(Z_i = z)$ has the Bernoulli distribution (which is a special case of the binomial distribution) with parameter $p_z = P(Z_i = z)$; $p_z = e^{-\nu} \nu^z / z!$. Therefore, Y_{ot} has the binomial distribution with parameters n and P_{t-d_0} . This distribution reflects the situation when a sample of fixed size n is taken from an infinite population where each subject has an equal and independent probability p_z of having an incubation period of z days:

$$p_z = e^{-\nu} \nu^z / z! = \nu / z (e^{-\nu} \nu^{z-1} / (z-1)!) \quad (4)$$

these assumptions form a framework for simulating time series of infectious processes with pre-specified properties.

As an example, Figure 1 shows the time series of daily counts simulated in the manner described above. Panels A and B illustrate a 70-day time series of a random process of infection with an outbreak initiated by an episode of exposure occurred on 24th day. For panel A, the mean endemic level was set to be 0.75 cases per day; the mean incubation time was 4.5 days; 15 cases were associated with the outbreak that starts on the 25th day. For a plot in the panel B the mean endemic level was set to be 1.75 cases per day; the mean incubation time was 7.5 days; and 45 cases were associated with an outbreak that also starts on the 25th day. In both simulated time series the duration of an outbreak is set to be three times the mean incubation period. The superimposed line is the daily mean used for simulations. This simulation imitates an outbreak with a single episode of high population exposure occurring on a pre-specified day without secondary spread of infection. A practical application of modeling single-source outbreaks of waterborne cryptosporidiosis using the reported were shown in Naumova (2000).

3. MODELING DISTRIBUTED EFFECTS OF EXPOSURE

Now we extend this framework to consider distributed effects of multiple episodes of exposure on the outcome of interest. We assume that the distributions of incubation periods in the population are similar for each episode of exposure to a specific pathogen. First, on any given day the observed counts of diseases, Y_t , consist of a number of cases caused by exposure occurring on different days: z_0 -cases occur due to exposure on a given day, z_1 -cases occur due to exposure on one day before, z_2 -cases occur due to exposure two days

before, etc. so $Y_t = \sum_{k=0}^{\infty} z_k$. Second, an exposure X_t occurring on a given day contributes to disease incidence for a number of days following the day of exposure, so that on any given day a variable $u_t = \sum_{k=0}^{\infty} \lambda_k X_{t-k}$ reflects the effect of exposure on a given day and is distributed over a time period according to a specific delay function that is characteristic to the causal pathogen. Let multipliers λ_k be Poisson-distributed and define the probability of an exposure to cause illness on day $i - k$ for those who were reported ill on day i which is $P(X_i = k) = e^{-v} v^k / k!$, where v is the population mean incubation period for the responsible pathogen. Considering a Poisson-distributed weight scheme:

$$\lambda_i = e^{-v} v^i / i!, \quad (5)$$

where $\sum_{i=0}^{\infty} \lambda_i = 1$, $0 \leq \lambda_i \leq 1$, $\lambda_0 = e^{-v}$, an iterative rule for the coefficients is

$$\lambda_i = \lambda_0 v^i / i! = \lambda_{i-1} v / i$$

$$\text{and } C = \sum_{i=0}^K \frac{v^i}{i!} e^{-v} = \sum_{i=0}^K w_i = 1.$$

For practical computational reasons, the maximum length of delayed effect, $\max(k) = K$, can be selected so $\sum_{i=0}^K w_i^*$ would be sufficiently close to 1 and cover the extent of an outbreak. In this case, K is a function of the mean incubation period and determines the limits on number of cases resulting from exposure.

To simulate a time series of infections with multiple episodes of exposures occurring on pre-specified days without secondary spread of infection, we generated a binary time series of high exposure events, X_t . In order to imitate the suggested lag structure empirically we introduced a lagged endogenous variable (Jorgensen, 1966), which represent a composite of distributed effects. We estimated the potential effect of exposure on a given day,

$u_t^* = \sum_{k=0}^{\infty} w_k^* X_{t-k}$, which then is used to simulate daily time series of cases attributable to an outbreak caused by exposure X_t . Now we let $D_t = 0$, if $u_t^* = 0$ and $D_t = 1$, if $u_t^* > 0$. Then the number of events on any given day t , is $Y_t = Y_{e_t} + D_t Y_{o_t}$. We then simulated daily cases as a composition of two Poisson random variables with respect to estimated curve u_t^* and number of cases attributable to each outbreak.

For example if we simulate a 70-day long time series that has a number of episodes of high exposure from a binomial distribution with parameter 0.25, the mean endemic level at 0.75 cases per day and the mean incubation time 4.5 days, then there are 15 episodes of exposure and 239 cases of infection with a daily mean of 3.14 ± 2.98 cases. This translates into 15 additional cases distributed over 14 days as defined in Equation (4).

To estimate the distributed effect of exposure on health response we can fit a regression model:

$$Y_t = \beta_0 + \beta_1 (\lambda_0 X_t + \lambda_1 X_{t-1} + \lambda_2 X_{t-2} + \dots + \lambda_i X_{t-i} + \dots + \lambda_K X_{t-K}) + \varepsilon \quad (6)$$

which is approximately equivalent to a linear model with a distributed lag structure that has the following form:

$$Y_t = \beta_0 + \beta_1 \sum_{i=0}^K w_i^* X_{t-i} + \varepsilon_t \quad (7)$$

Considering a Poisson-distributed weight scheme as (4) corrected for a fixed K , a linear model with a Poisson-distributed lag structure and Gaussian assumptions for the error term can be re-written as follows (co-called in the text as Gaussian model):

$$Y_t = \beta_0 + \beta_1 u_t^* + \varepsilon \quad (8)$$

where β_0 is an estimate of endemic level of infection, β_1 an estimate of overall effect of time-distributed exposure, and $u_t^* = \sum_{i=0}^K w_i^* X_{t-i}$ is a lagged endogenous variable. The model (8) can be modified to allow a positive discrete outcome variable by using a Poisson distribution for the error term and a log-link function (co-called in the text as Poisson model):

$$Y_t = \exp\{\beta_0 + \beta_1 u_t^* + \varepsilon_t\} \quad (9)$$

Both models (8) and (9) can include additional variables, V_t :

$$Y_t = \beta_0 + \beta_1 u_t^* + \beta_2 V_t + \varepsilon_t \quad (10)$$

and

$$Y_t = \exp\{\beta_0 + \beta_1 u_t^* + \beta_2 V_t + \varepsilon_t\} \quad (11)$$

To illustrate the performance of Gaussian and Poisson models, as defined in (8) and (9), respectively, we fitted both models to the simulated time series and examined the recovery of the simulated mean. The results of modeling are given in Table 1. Both models demonstrated good performance, although the Gaussian model fitted the data better. One of the reasons for the observed advantage of the Gaussian model is that in the presence of multiple closely spaced outbreaks the outcome distribution could be viewed as a mixture of Poisson distributions, which is better approximated by a normal distribution.

Based on values of regression parameters we estimated the mean endemic level, as $\beta_0^* \pm se_\beta$ for the Gaussian model and as $\exp(\beta_0^* \pm se_\beta)$ for the Poisson model using the intercept and its standard error estimates. The total number of predicted cases $\sum_{t=1}^N Y_t^* = \sum_{t=1}^N Y_{E_t}^* + \sum_{t=1}^N Y_{O_t}^*$, where the first term reflects the total number of cases attributed to endemic level and the second term reflects the total number of cases attributed to all episodes of high exposure, can be estimated for the Gaussian model as

$$\sum_{t=1}^N Y_t^* = \beta_0^* N + \beta_1^* \sum_{t=1}^N u_t^* \quad (12)$$

and for the Poisson model as

$$\sum_{t=1}^N Y_t^* = \sum_{t=1}^N \exp\{\beta_0^* + \beta_1^* u_t^*\} \quad (13)$$

In our simulated example, the Poisson model provides the estimates for an overall mean endemic level of 1.69 ± 1.13 cases per day and 13.74 cases associated with an episode of high exposure. The Gaussian model offers somewhat smaller values: 0.95 ± 0.44 cases for the endemic level and 13.5 cases associated with an episode of high exposure. As expected, due to multiple episodes of exposure, the estimates for the endemic level are higher than the simulated endemic rate of 0.75 cases per day, but much smaller than the sample average of 3.6 endemic and epidemic cases per day. As we gain experience with the distributed lag model, an inflator correcting for frequency of exposure can be incorporated into the modeling procedure.

4. ESTIMATION OF THE MEAN INCUBATION PERIOD

In the example presented above, we assumed that the mean incubation period is known. In practice, often only the range of incubation times is available for a specific pathogen. By assuming a Poisson distribution of incubation periods in the population, we developed an approach to estimating the mean incubation period from an actual time series of observed cases. If an epidemic curve follows a Poisson structure as in Equation (4), then the time-distributed effect of exposure for a sample of fixed size, n (taken from an infinite population where each subject has an equal and independent probability p_z of having an incubation period of z days), starts at day d_0 . The daily mean incidence for two consecutive outbreak days, s_i and s_{i+1} , has the following relations: $s_{i+1}/s_i = v/i$, where i is the number of days after

the start of an outbreak and v is the population mean incubation period for the responsible pathogen. For practical consideration summing over the length of incubation period, K , limits the number of cases resulting from an episode of exposure. Therefore, the mean incubation period can be roughly estimated as follows:

$$\widehat{v} = \frac{1}{K+1} \sum_{i=0}^K i s_i / s_{i+1} \quad (14)$$

Misspecification of the mean incubation period could be a potential problem for distributed lag models. We demonstrated the effect of lag misspecification by contrasting both models. We simulated exposure and outcome variables as described above, except the 70-day long time series had only two episodes of high exposure. Each episode of high exposure caused 15 additional cases distributed over 14 days; the mean endemic level was 0.75 cases per day; the mean incubation time was 4 days. We repeated simulations 1000 times (an example of a simulated time series is shown in Figure 2A). To each simulated time series we fitted 12 regression models as defined by formula (9) to demonstrate the fit of a pre-selected lag that varied from 1 to 12 days. As the mean lag progress from 1 to 12, the quality of fit continually changed. As expected, models exhibited the best fit at the mean distributed lag of four days (see Figure 2B). An iterative procedure of this kind can be performed as an exploratory step to examine the potential range of incubation times.

5. MODEL SENSITIVITY

Model fit may potentially depend on the frequency of high exposure episodes, length of time series, average incubation time, effect size, and other factors. To study such dependencies we examine model fit at different levels of effect size and over the entire range of frequency of exposure episodes. We also examine model sensitivity to the population mean incubation period of the disease. First, we developed an iterative simulation scheme to reflect the occurrences of high exposure episodes. At each of 100 replications, m , we simulated a 1000-day long time series, X_{ijm} , where $t = 1, \dots, 1000$, $m = 1, \dots, 100$ using a generator of pseudo-random numbers. We draw realizations of X_{ijm} from a binomial distribution with parameter a , where $0 < a < 1$, $a = 0.05j$, $j = 1, \dots, 19$. Next, for each simulated set X_{ij} , we generated a 1000-day long time series for the outcome variable with the mean endemic level of one case per day, and with L episodes of high exposure, where $L = 1000a$. Assuming that each episode of high exposure initiates an outbreak with S_L cases distributed over a period of time that is three times the length of the mean incubation time, we set a magnitude of a single outbreak of $S_L = 5n$, $n = 1, \dots, 10$ (Figure 3).

In these outbreaks, cases followed a Poisson distribution with the mean incubation time of 1.5 days (Figure 3: panels A and D). We repeated simulations for incubation periods of 4.5 days (Figure 3: panels B and E) and 7.5 days (Figure 3: panels C and F). Thus, we performed $10 \times 19 \times 100$ simulations for each of three mean incubation periods. The results are shown as an image plot where the x -axis represents the fraction of days with high exposure, the y -axis reflects the magnitude of outbreaks initiated by the each episode of exposure, and the z -axis indicates percent of variability explained by the model. The color gradient reflects progression from poor fit of less than 15% (dark blue), to a moderate fit (yellow–orange), and to an excellent fit of more than 85% recovery (light pink). As expected, a better model fit was achieved when the fraction of days with high exposure is low, the magnitude of outbreaks initiated by each episode of exposure is high, and the incubation period is short. The Gaussian model fits slightly better than the Poisson model at low frequency of exposure; the Poisson model out-performs the Gaussian model if in a 1000-day time series the number of episodes of exposure ranges between 200 and 600.

Overall, the percent of variability explained by Gaussian and Poisson models correlates very well ($r = 0.975\text{--}0.989$).

6. HIGH AMBIENT TEMPERATURE AND CRYPTOSPORIDIOSIS

The hypothesis for the presented analysis is that hot days in the summer are associated with an increase in the incidence of waterborne cryptosporidiosis. To test this hypothesis we decomposed a time series of daily counts of laboratory-confirmed cases into systematic and random components considering presence of a linear trend, periodic seasonal fluctuations, and effect of the day of the week and holidays.

The 10-year time series of daily ambient temperature and reported cases of *Cryptosporidium* infections (1992–2001) are shown in Figure 4A and B. The time series of infections were created using dates of onset for laboratory-confirmed cases reported in the state of Massachusetts. Daily ambient temperature values were determined as an average of daily maximum temperature from all 38 active stations in MA. We define a high ambient temperature as daily average above 90°F, but also explored two other cut points of 85 and 80°F. The graph of reported cryptosporidiosis cases shows three distinct time periods. The first period has low reported incidence and spans from the initiation of the surveillance in 1992 until late summer of 1995 when the outbreak occurred in Worcester, MA (Naumova *et al.*, 2000). The second period can be associated with peak incidence of reported cryptosporidiosis during the outbreak and the third period—with increased reporting after the outbreak.

In 10 years, there were 34 days when the average temperature exceeded 90°F. During the post-outbreak period, when the fraction of hot days was less than one percent, a spike of five cases which occurred in the summer of 1999 followed a well-defined heat wave period. Seasonal fluctuations in the incidence of *Cryptosporidium* infections were very prominent with increased reporting in the late August; the potential for high incidence in cryptosporidiosis in early spring was also noted elsewhere (Naumova *et al.*, 2005a).

For this analysis, we used only the post-outbreak data (1996–2001) with the mean daily incidence of 0.16 ± 0.57 cases. The distribution of daily counts of infections was skewed to the right. We included two harmonic components to adjust for seasonality (Naumova and MacNeill, 2006), a linear component for the trend and a set of indicator variables to adjust for the day of the week effect and for the first day of the months (occasionally used by the reporting system if the actual calendar day was missing). To reflect occurrences of temperature $> 90^\circ\text{F}$ we created a binary variable. We also identified eleven federal holidays and designed a set of indicator variables for each week following the holiday. Table 2 indicates the number of days in the studied time series marked as associated with one of 11 holidays.

To examine a distributed effect of high temperature with the average delay ranging from one to ninety days, we fit the following regression model, which is analogous to (11):

$$Y_t = \exp\{\beta_0 + \beta_1 u_t + \beta_2 t + \beta_3 \cos(2\pi\omega t) + \beta_4 \sin(2\pi\omega t) + \beta_5 \cos(4\pi\omega t) + \beta_6 \sin(4\pi\omega t) + \beta_7 H + \beta_8 D + \beta_9 F + \varepsilon\} \quad (15)$$

where u_t is a variable designed to reflect a distributed effect of exposure occurring proportionally to a specific delay function characteristic of a pathogen for a pre-selected exposure, the third term indicated linear trend, the next four terms account for seasonality, the next three terms account for holidays, effect of the week, and reporting defaults. The

mean lag was set to be varying from one day to 13 days to cover a range of plausible incubation periods of the illness.

Results suggest that the distributed lag model with the mean lag of seven days (see figure 5) fitted the cryptosporidiosis data the best, yielding the value for the regression parameter of 90°F 0.53777188 (0.03637577) and suggesting that 7.22 cases distributed over ~21 days might be attributed to an episode of hot weather. The mean lag of seven days is consistent with the typical mean incubation period for cryptosporidiosis.

7. DISCUSSION

In this communication we demonstrated an application of a model for time-distributed effects of exposures to infection cases with respect to an incubation period of infection. The model provided a realistic estimate of cases of waterborne cryptosporidiosis distributed over relatively short period and attributed to an episode of potentially high exposure. This study was motivated by various observations suggesting that meteorological factors may drive seasonal patterns in probability of exposure and transmission for waterborne pathogens, and fluctuations in the incidence of enteric infections on a short temporal scale may be caused by social and environmental factors influenced by an extreme precipitation event (Curriero *et al.*, 2001; Naumova *et al.*, 2005c) or ambient temperature (Bentham and Langford, 2001). However, the analysis of the impact of environmental temperature on temporal variations in enteric infections is limited due to insufficient analytical tools available to public health professionals.

We demonstrated that the model fit may depend on a number of factors: frequency of high exposure episodes, estimate of incubation time, size of anticipated effect, and other factors. Such dependencies could be nonlinear and deserved further exploration. One of the important issues to be addressed is a potential for serial autocorrelation that can be still left in the residuals. Examining actual data on reported cases of cryptosporidiosis obtained from surveillance monitoring in Massachusetts, USA, we found a tendency in this data to be highly skewed. Further improvement in estimation of model parameters can be achieved by applying a model with negative-binomial distributional assumptions. Further development on heavy-tail modeling associated the time-distributed effects of exposure may find a new area of application in studying infectious outbreaks.

To gain more epidemiological insight into pathogen transmission and population susceptibility, the suggested modeling can be further expanded to consider different temperature cut-points tailored for specific climatic conditions, different subpopulations (e.g., different age group or populations served by different water sources), and different infections. Systematic temporal fluctuations in diseases that have similar sources of exposure, routes of transmissions, and clinical manifestations might imply a presence of environmental factors that synchronize temporal increases and form seasonal curves. The proposed model may find various applications in assessing the effects of environment on human health.

Acknowledgments

We wish to thank Dr. Jeffrey Griffiths, Dr. Andrey Egorov, and our anonymous reviewers for their thoughtful suggestions, and also Dr. Alfred DeMaria, Dr. Bela Matyas and the Massachusetts Department of Public Health for providing us with surveillance data. We would also like to thank agencies that funded this study: the USA EPA (68-C-02-042), the National Institute of Allergy and Infectious Diseases (AI43415, U19A1062627, HHSN266200500032C), and the National Institute of Environmental Health Sciences (ES013171).

References

- Almon S. The distributed lag between capital appropriations and expenditures. *Econometrica*. 1965; 33:178–196.
- Barwick RS, Levy DA, Craun GF, Beach MJ, Calderon RL. Surveillance for waterborne-disease outbreaks—United States, 1997–1998. *Morbidity and Mortality Weekly Report. CDC Surveillance Summaries*. 2000; 49(4):1–21. [PubMed: 10843502]
- Bentham G, Langford IH. Climate change and the incidence of food poisoning in England and Wales. *International Journal of Biometeorology*. 1995; 39(2):81–86. [PubMed: 8530209]
- Bentham G, Langford IH. Environmental temperatures and the incidence of food poisoning in England and Wales. *International Journal of Biometeorology*. 2001; 45(1):22–26. [PubMed: 11411411]
- Brock, F. Victims of Isolation. Vol. Section 3. *New York Times*; September 14. 2003 Seniority: Victims of the Heat; p. 8
<http://query.nytimes.com/gst/abstract.html?res=F50D11F63A5F0C778DDDA00894DB404482>
- Chetty VC. Estimation of Solow's distributed lag models. *Econometrica*. 1971; 39(1):99–117.
- Cowden JM, Wall PG, Adak G, Evans H, Le Baigue S, Ross D. Outbreaks of foodborne infectious intestinal disease in England and Wales: 1992 and 1993. *Communicable Disease Report. CDR Review*. 1995; 5(8):R109–R117. [PubMed: 7663603]
- Curriero FC, Patz JA, Rose JB, Lele S. The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948–1994. *American Journal of Public Health*. 2001; 91(8):1194–1199. [PubMed: 11499103]
- Djuretic T, Wall PG, Nichols G. General outbreaks of infectious intestinal disease associated with milk and dairy products in England and Wales: 1992 to 1996. *Communicable Disease Report. CDR Review*. 1997; 7(3):R41–R45. [PubMed: 9080728]
- Egorov A, Naumova E, Tereschenko A, Kislitsin V, Ford T. Daily variations in effluent water turbidity and diarrheal illness in a Russian city. *The International Journal of Environmental Health Research*. 2003; 13:81–94.
- Evans HS, Madden P, Douglas C, Adak GK, O'Brien SJ, Djuretic T, Wall PG, Stanwell-Smith R. General outbreaks of infectious intestinal disease in England and Wales: 1995 and 1996. *Communicable Disease and Public Health*. 1998; 1(3):165–171. [PubMed: 9782630]
- Fleming CA, Caron D, Gunn JE, Horine MS, Matyas BT, Barry MA. An outbreak of *Shigella sonnei* associated with a recreational spray fountain. *American Journal of Public Health*. 2000; 90(10):1641–1642. [PubMed: 11030007]
- Fuchs, D. World Briefing: Eurpoe: Spain: Deaths Rose in Heat Wave. Vol. Section 3. *New York Times*; September 14. 2003 p. 8
<http://query.nytimes.com/gst/fullpage.html?res=9B04E7D61538F931A3575AC0A9659C8B63>
- Hobbs, BC.; Roberts, D. *Food Poisoning and Food Hygiene*. Edward Arnold; London: 1987.
- Huang, Y.; Dominici, F.; Bell, ML. Bayesian Hierarchical Distributed Lag Models for Summer Ozone Exposure and Cardio-Respiratory Mortality. Johns Hopkins University, Department of Biostatistics Working Papers; October 31. 2004 Working Paper 56
- Joce RE, Bruce J, Kiely D, Noah ND, Dempster WB, Stalker R, Gumsley P, Chapman PA, Norman P, Watkins J. An outbreak of cryptosporidiosis associated with a swimming pool. *Epidemiology and infection*. 1991 Dec; 107(3):497–508. [PubMed: 1752299]
- Jorgensen DW. Rational distributed lag functions. *Econometrica*. 1966; 32(1):135–149.
- Kapperud G, Skjerve E, Bean NH, Ostroff SM, Lassen J. Risk factors for sporadic campylobacter infections: results of a case-control study in southeastern Norway. *Journal of Clinical Microbiology*. 1992; 30(12):3117–3121. [PubMed: 1452694]
- Koyck, LM. *Distributed Lags Models and Investment Analysis*. North-Holland; Amsterdam: 1954.
- Kusumaningrum HD, Riboldi G, Hazeleger WC, Beumer RR. Survival of foodborne pathogens on stainless steel surfaces and cross-contamination to foods. *International Journal of Food Microbiology*. 2003; 85(3):227–236. [PubMed: 12878381]
- Lee SH, Levy DA, Craun GF, Beach MJ, Calderon RL. Surveillance for waterborne-disease outbreaks—United States, 1999–2000. *Morbidity and Mortality Weekly Report. Surveillance Summaries*. 2002; 51(8):1–47. [PubMed: 12489843]

- Morris RD, Naumova EN, Griffiths JK. Did Milwaukee Experience Waterborne Cryptosporidiosis Before the Large Documented Outbreak in 1993? *Epidemiology*. 1998; 9(3):264–270. [PubMed: 9583417]
- Naumova, EN. Modeling a signature of a single-exposure waterborne outbreak with secondary spread. In the Proceedings of the Joint Statistical Meetings. Section: Biometrics; 2000. p. 76-81.
- Naumova EN, MacNeill IB. Signature-forecasting and early outbreak detection system. *Environmetrics*. 2005; 16:749–766. [PubMed: 18716671]
- Naumova, EN.; MacNeill, IB. Seasonality assessment for biosurveillance systems. In: Balakrishnan, N.; Auget, J.-L.; Mesbah, M.; Molenberg, G., editors. *Advances in Statistical Methods for the Health Sciences: Applications to Cancer and AIDS Studies, Genome Sequence Analysis, and Survival Analysis*. Birkhauser; Boston: 2006.
- Naumova, EN.; Terrin, N.; Morris, RD. A quantitative framework for characterizing waterborne outbreaks. In the Proceedings of the JASA Meeting; Baltimore, USA. August 1999; p. 117-122.
- Naumova EN, Chen JT, Griffiths JK, Matyas BT, Estes-Smargiassi S, Morris RD. The use of passive surveillance data to study temporal and spatial variation in the incidence of giardiasis and cryptosporidiosis. *Public Health Reports*. 2000; 151:38–50.
- Naumova EN, Egorov AI, Morris RD, Griffiths JK. The elderly and waterborne *Cryptosporidium* infection: gastroenteritis hospitalizations before and during the 1993 Milwaukee outbreak. *Emerging Infectious Diseases*. 2003; 9(4):418–425. [PubMed: 12702220]
- Naumova EN, Christodouleas J, Hunter PR, Sued Q. Temporal and spatial variability in cryptosporidiosis recorded by the surveillance system in North West England in 1990–1999. *Water and Health*. 2005; 3(2):185–196.
- Naumova EN, O’Neil EA, MacNeill IB. INFERNO: a system for early outbreak detection and signature forecasting. *MMRW Supplement*. 2005; 54(Suppl):77–83.
- Naumova EN, Christodouleas J, Hunter PR, Sued Q. Temporal and spatial variability in cryptosporidiosis recorded by the surveillance system in North West England in 1990–1999. *Water and Health*. 2005; 3(2):185–196.
- Naumova EN, Jagai J, Matyas B, DeMaria A, MacNeill IB, Griffiths JK. Enteric diseases seasonality and ambient temperature in Massachusetts. *Epidemiology and Infections*. 2006; 135(2):281–292.
- Neimann J, Engberg J, Molbak K, Wegener HC. A case-control study of risk factors for sporadic campylobacter infections in Denmark. *Epidemiology and Infection*. 2003; 130(3):353–366. [PubMed: 12825719]
- Schwartz J. The distributed lag between air pollution and daily death. *Epidemiology*. 2000; 11:320–326. [PubMed: 10784251]
- Sockett PN. Food poisoning outbreaks associated with manufactured foods in England and Wales: 1980–1989. *CDR (London: England Review)*. 1991 Sep 13; 1(10):R105–R110.
- Solow R. On a family of lag distributions. *Econometrica*. 1960; 28:393–406.
- Welty, LJ.; Zeger, SL. Johns Hopkins University, Department of Biostatistics Working Papers; 2005. Are the Acute Effects of PM10 on Mortality in NMMAPS the Result of Inadequate Control for Weather and Season? A Sensitivity Analysis Using Flexible Distributed Lag Models (January 25, 2005). Working Paper 38. <http://www.bepress.com/jhubiostat/paper38>
- Wyzga RE. The effect of air pollution upon mortality: a consideration of distributed lag models. *Journal of the American Statistics Association*. 1978; 73(363):463–472.

Figure 1.

Simulated 70-day long time series for a time-distributed response (events) initiated by an episode of exposure occurred on 24th day. Panel A: the mean endemic level—0.75 cases per day; the mean incubation time—4.5 days; outbreak magnitude—15 cases. Panel B: the mean endemic level—1.75 cases per day; the mean incubation time—7.5 days; outbreak magnitude—45 cases. In both scenarios, the duration is 3 times of the mean incubation period; the superimposed line is the daily mean of counts of events

Figure 2.

Simulated 70-day long time series of daily counts of disease with mean endemic level of 0.75 cases per day and two simulated outbreaks occurred 15 days apart. Each outbreak has mean incubation time of 4 days associated with 15 additional infection cases distributed over 14 days (Panel A). The superimposed line is the daily mean of counts of events. In Panel B: estimates of β_1 from fitted models with the mean lag ranging from 1 to 12 days. The distributed lag model with the mean lag of 4 days fitted the simulated data the best

Figure 3.

Image plots of the per cent of variability explained by the Gaussian (Panels A, B, and C) and Poisson models (Panels D, E, and F) as the function of the frequency of exposure based on the fraction of days with high exposure (x -axis) and the magnitude of outbreaks initiated by the each episode of exposure (y -axis) for three values of mean incubation time (MIT): 1.5 days (Panels A and D), 4.5 (Panels B and E), and 7.5 days (Panels C and F). The color gradient from dark blue through hues of green, yellow, and brown, to light pink reflects the quality of fit ranged from 0.1 to 85%, respectively. Image plots of the per cent of variability explained by the Gaussian model as the function of the fraction of days with high exposure (x -axis) and the magnitude of outbreaks initiated by the each episode of exposure (y -axis) for three values of MIT: 1.5, 4.5, and 7.5 days

Figure 4. Time series of daily ambient temperature with three pre-selected levels: 80, 85, 90°F (panel A); Time series of daily counts of cryptosporidiosis (panel B) with marked days of ambient temperature above 90°F (vertical dotted lines)

Figure 5.

Estimates of the regression parameter in the distributed lag model with the mean lag of 1–12 days fitted the cryptosporidiosis time series (post 1995 outbreak). The distributed lag model with the mean lag of 7 days fitted the cryptosporidiosis data the best, yielding the value for the regression parameter of 0.481 (0.230) and suggesting that 7.22 cases distributed over ~21 days might be attributed to an episode of hot weather

Table 1

The results of modeling of time-distributed effect using Poisson and Gaussian assumptions in an outcome distribution for simulated time series

Parameter estimates	Value	Std. error	T-value	Variance explained
Poisson model				0.5055692
Intercept, β_0^*	0.5065428	0.1214127	4.172073	
Exposure, β_1^*	3.1270396	0.3123925	10.009969	
Gaussian model				0.5343453
Intercept, β_0^*	0.947716	0.4439239	2.134861	
Exposure, β_1^*	13.499472	1.5282110	8.833513	

Table 2

Number of days in holiday weeks in the 10-year time series

	Holiday	Days
New year's	January 1	81
MLK	3rd Monday in January	74
Superbowl	Last Sunday in January	80
President's	3rd Monday in February	80
Patriot's	3rd Monday in April	80
Memorial	Last Monday in May	80
Independence	July 4	83
Labor	1st Monday in September	80
Columbus	2nd Monday in October	80
Thanksgiving	4th Thursday in November	90
Christmas	December 25	72
Days with no holidays		2773