ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2015)
178, *Part* 1, *pp.* 205–222

Modelling reporting delays for outbreak detection in infectious disease data

Angela Noufaily, Yonas Ghebremichael-Weldeselassie, Doyo Gragn Enki and Paul Garthwaite,

The Open University, Milton Keynes, UK

Nick Andrews and André Charlett

Public Health England, London, UK

and Paddy Farrington

The Open University, Milton Keynes, UK

[Received June 2013. Final revision November 2013]

Summary. The delay that necessarily occurs between the emergence of symptoms and the identification of the cause of those symptoms affects the timeliness of detection of emerging outbreaks of infectious diseases, and hence the ability to take preventive action. We study the delays that are associated with the collection of laboratory surveillance data in England, Wales and Northern Ireland, using 12 infections of contrasting characteristics. We use a continuous time spline-based model for the hazard of the delay distribution, along with an associated proportional hazards model. The delay distributions are found to have extremely long tails, the hazard at longer delays being roughly constant, suggestive of a memoryless process, though some laboratories appear to stop reporting after a certain delay. The hazards are found typically to vary strongly with calendar time, and to a lesser extent with season and recent organism frequency. In consequence, the delay distributions cannot be assumed to be stationary. These findings will inform the development of outbreak detection algorithms that take account of reporting delays.

Keywords: Delay; Hazard; Infectious disease; Penalized likelihood; Spline; Surveillance

1. Introduction

Delays are ubiquitous in surveillance data since only rarely, if ever, can a health event of interest be recognized or registered instantaneously. Delays may be intrinsic to the underlying biological process of interest, as with the incubation period of an infection, namely the time elapsed between an individual's being infected and the appearance of symptoms. Alternatively, delays may involve external processes, notably the time taken to reach a diagnosis, or to enter this diagnosis in the appropriate database, as is the case with cancers.

The statistical analysis of delay data has a history stretching back at least 50 years. An early reference is Sartwell's study of incubation periods using log-normal distributions (Sartwell, 1966). More recently, Bayesian methods have been used in this context to handle situations in

Address for correspondence: Angela Noufaily, Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA,UK.
E-mail: Angela.Noufaily@open.ac.uk

 0964–1998/15/178205

which infection times are unobserved (O'Neill *et al.*, 2000). However, the major application of statistical methods to the analysis of delay data in disease surveillance has been to obtain valid estimates of recent incidence, when case ascertainment or reporting is subject to delays. Thus, the advent of the acquired immune deficiency syndrome epidemic in the 1980s, and the urgency with which estimates were required of current trends in the incidence of human immunodeficiency virus, provided a great impetus to the statistical treatment of delay data (Brookmeyer and Gail, 1988; Lawless, 1994).

Methods for estimating delay distributions from grouped data were first described by Harris (1989), Brookmeyer and Damiano (1989) and Zeger *et al.* (1989), within a more general formulation involving joint estimation of the incidence function as well as the delay distribution. Brookmeyer and Liao (1990) proposed a convenient method for the analysis of grouped delay data that does not involve the incidence function. This model, which can be implemented as a generalized linear model (McCullagh and Nelder, 1989), has been used to study factors influencing delays in acquired immune deficiency syndrome epidemiology, notably whether delays are changing over time (Bacchetti, 1996; Cui and Kaldor, 1998; Gebhardt *et al.*, 1998; Deuffic and Costagliola, 1999; Tabnak *et al.*, 2000). Recently, the model has been used to model reporting delays for salmonellas (Jones *et al.*, 2014). The model is particularly convenient when implemented with a complementary log–log-link, in which case it induces a relationship between the cumulative distribution functions (CDFs) of the delay distribution of the form

$$F(t|x) = F(t)^{\exp(\beta^{\mathrm{T}} x)} \tag{1}$$

where $F(t|x)$ is the CDF of delays in a subpopulation with characteristics x and $F(t) = F(t|0)$. Brookmeyer and Liao (1990) noted that this model is not a proportional hazards model: when $\beta^{\mathrm{T}} x > 0$ the ratio of hazard functions increases monotonically from 0 to 1, thus converging at long delays. The model has been described as a 'reverse time proportional hazards model' (Kalbfleisch and Lawless, 1989).

Midthune *et al.* (2005) generalized the model of Harris (1989) to include corrections to the data as arise in cancer registries. The reporting delay component of the model, which as its predecessors is specified for grouped data, parameterizes the probability of a given delay in terms of the discrete hazard, using a complementary log–log-link function, though the model is not a generalized linear model.

The methods that are developed in the present paper are motivated by the LabBase laboratory surveillance data which are routinely collected on a large number of distinct organism types by Public Health England. These data are used to detect, monitor and if appropriate to take control measures to stem the development of outbreaks (Enki *et al.*, 2013).

In this paper we do not address in any detail how reporting delays ought to be taken into account in outbreak detection. There, the focus is wholly on reports with relatively short delays, since those with long delays cannot be acted on promptly. However, it is also of interest to examine the entire reporting delay distribution, to obtain a better understanding of the reporting process, and perhaps to help to improve it. Thus, our aim in this paper is to describe the reporting delay hazard, with the aim of gaining a better understanding of the reporting process, and to investigate whether temporal factors, notably calendar time, season and recent incidence, influence reporting delays. Similar considerations have recently motivated Jones *et al.* (2014) who studied in detail the factors affecting reporting delays for salmonellas in France. Two features distinguish the UK data from most other applications: the great diversity of organisms involved, and the very long tails of the delay distributions. To capture these features in a unified framework, we have adopted a proportional hazards modelling approach in continuous time, the baseline hazard being represented by spline functions.

In the next section, we describe the data and motivate our modelling approach. Then in Section 3 we derive the likelihood and describe the spline model for the hazard. Section 4 contains the data analysis.

2. Reporting delay data for infectious diseases

2.1. The LabBase system

In England, Wales and Northern Ireland, several hundred hospital and specialist laboratories send individual reports of isolates of infectious pathogens—usually typed for epidemiological purposes down to a detailed classification such as serotype or phage type—to the central LabBase database at Public Health England (previously the Health Protection Agency) in London. These data have been collected electronically since 1991 and comprise several thousand different organism types (Enki *et al*., 2013). They constitute a key resource in infectious disease surveillance; an automatic outbreak detection algorithm is run each week to detect potential outbreaks (Farrington *et al*., 1996; Noufaily *et al*., 2013). This algorithm seeks to identify organisms for which the current weekly count is unusually high relative to the expected count. Such aberrant organisms are then subjected to further scrutiny.

Each reported isolate contains the date at which the specimen (which might be a sample of blood, urine, faeces etc.) was taken, and the date at which the report was sent to LabBase. The time interval between the specimen date and the report date is the reporting delay. This includes the time that is needed to analyse and classify the specimen but may also include clerical and other delays. Data entry errors in the dates may also affect the reporting delays. The LabBase outbreak detection system is based on dates of report; some of the issues relating to this are discussed in Farrington *et al*. (1996). It is hoped that the present paper will help to inform a move to outbreak detection based on dates of specimen.

Reporting delays are known to differ substantially between infectious pathogens (which require different identification procedures, of varying complexity) and between reporting laboratories (some of which are reference laboratories specializing in certain types of organisms). Some specimens are sent to several laboratories in succession to undertake different typing steps. This applies, in particular, to salmonellas, for which local laboratories undertake a partial identification, after which isolates are sent to the national reference laboratory for full characterization. Reporting delays may also vary according to season, or according to the current workload, or over time. The aim of the present paper is to investigate these types of temporal variation. When using surveillance data for outbreak detection purposes, time is of the essence. Thus there is little time to check and correct the data, other than excluding data with obvious errors (such as reports entered as occurring before specimen collection). The data that are analysed here are raw data, as used in practice for outbreak detection.

2.2. Description of the data

We report on 12 different infectious organisms, of contrasting characteristics. These organisms were selected to provide a wide range of organism types (bacteria, viruses and protozoa, with different transmission routes), seasonalities, median frequencies and reporting delays. Seven salmonella serotypes were included, in view of the public health importance of salmonella infection. The data were provided by Public Health England from their LabBase surveillance database. All isolates with dates of report between January 1st, 2004, and December 31st, 2011, were obtained for these 12 organisms. The date of specimen and date of report were used to calculate the reporting delay, in days. Information on reporting laboratories was not available.

Table 1. The 12 organisms analysed†

Organism name	*N*	N_1 *(%)*	N_2 *(%)*
Acinetobacter baumanii	4033	6 (0.15)	5 (0.12)
Campylobacter jejuni	1984	6 (0.30)	1 (0.05)
Chlamydia sp	31127	31 (0.10)	338 (1.09)
Giardia lamblia	27243	48 (0.18)	29 (0.11)
Norovirus	56574	6 (0.01)	199 (0.35)
Salmonella abony	106	1 (0.94)	0 (0.00)
Salmonella braenderup	848	2 (0.24)	0 (0.00)
Salmonella brandenburg	177	0 (0.00)	1 (0.57)
Salmonella enteritidis PT21	3088	7 (0.23)	3 (0.10)
Salmonella infantis	1146	0 (0.00)	0 (0.00)
Salmonella senftenberg	480	1 (0.21)	1 (0.21)
Salonella typhimurium DT104	2373	5 (0.21)	6 (0.25)

†N is the total number of isolates, N_1 the number with negative delays and N_2 the number with delay greater than 730 days.

Table 1 lists the 12 organisms along with the total counts of isolates for each organism (N). Also shown are the numbers and percentages of reports for which the reporting delay is allegedly negative (N_1), i.e. the date of report precedes the date of specimen, presumably owing to data entry error. These isolates were excluded from further analysis. The earliest date of specimen was April 14th, 1994, the reporting delay being over 10 years. Such extreme intervals are likely also to be due to data entry error. Few isolates were found to have reporting delays greater than 730 days (numbering N_2 in Table 1): only for *Chlamydia* sp (where 'sp' denotes 'species') do these account for more than 1% of reports (328 of the 338 with delays over 2 years for *Chlamydia* sp having specimen dates before January 1st, 2002). As delays over 730 days are so infrequent, and appear to relate primarily to historical data, we excluded such reports. In fact, as will become apparent later, for several organisms there is evidence that isolates with long delays are systematically discarded by many laboratories.

Table 2 gives some brief indications about the clinical and epidemiological characteristics of the organisms that were selected, and Table 3 shows some descriptive statistics. For several organisms the minimum delay is 0 days, implying that the specimen was collected and the pathogen identified, typed and reported on the same day; however, only 10 delays among the many tens of thousands analysed were equal to 0, so no special measures were taken to account for these. The median delay varies from 9 to 25 days, and in all cases the mean is very much greater than the median, implying substantial right skew. The maximum values are all very high, indicating very long upper tails. In contrast, the 75% quantiles are in most instances of the order of a month or less.

2.3. Analysis strategy

We decided to model delays in continuous time, thus avoiding the need to classify delays into discrete categories which may need to differ between organisms. A further reason for using a continuous time framework is to visualize the hazard function over the span of the data and thus to gain a better understanding of the processes generating the delays. Finally, we decided to eschew the reverse time proportional hazards regression model in favour of an explicit model for the hazard, using splines to allow a flexible representation of hazards at moderately long delays, which are a striking feature of these data. The regression model that was used is a

Table 2. Clinical and epidemiological information on the 12 organisms

Organism name	*Clinical and epidemiological information*
Acinetobacter baumanii	Causes potentially serious bacteraemia, respiratory and wound infections; transmission by direct contact or contact with contaminated surfaces, notably in hospitals; peaks in summer
Campylobacter jejuni	Causes infectious intestinal disease; transmitted by consumption of contaminated food, contact with infected animals or person to person
Chlamydia sp	The most common sexually transmitted infection; no clear seasonality
Giardia lamblia	A waterborne protozoan infection causing diarrhoeal disease; transmitted by the faecal–oral route through contact with contaminated water or person to person; peaks in summer
Norovirus	A common cause of infectious gastroenteritis; highly infectious; transmitted from person to person; peaks in winter
Salmonellas	A common cause of food poisoning; most of the several thousand strains are transmitted by consumption of contaminated food or person to person; most salmonellas peak in summer

proportional hazards model providing a simple interpretation of parameters in terms of relative hazards. Note that inferences on covariate effects are driven by the shorter delays, which are much more numerous than longer delays. This is appropriate for our intended application to outbreak detection, where only short delays are of interest. In practice, we would not expect hazards to be proportional over the entire range of delays; however, the proportional hazards model provides a useful test of the null hypothesis of equal hazards, and the validity of the model can readily be investigated graphically.

3. Modelling reporting delays

3.1. Constructing the log-likelihood

Let the random variables S, T and D represent respectively the date of specimen, the date of report and the reporting delay for a particular organism. Thus $T \geqslant S$, $D = T - S$ and $D \geqslant 0$. Suppose that the probability density function of D is $f(\cdot)$, with CDF $F(\cdot)$.

In specifying the likelihood, we need to take into account both left and right truncation of the data. This truncation is a consequence of the fact that the data that we obtained relate to reports received between January 1st, 2004, and December 31st, 2011. Since delays are no greater than 730 days, for simplicity, we shall count days from day 1 corresponding to January 1st, 2002. Let $\tau_1 = 731$, corresponding to January 1st, 2004 (the first possible value of T), $\tau_2 = 2922$, corresponding to December 31st, 2009 (2 years before the last possible value of T), and $\tau_3 = 3652$, corresponding to December 31st, 2011 (the last possible value of T).

The contribution of an observation (S, T), with $D = T - S$, then depends on whether $S \leqslant \tau_1$, $\tau_1 < S \leqslant \tau_2$ or $\tau_2 < S \leqslant \tau_3$, via the appropriate conditioning to take into account the truncation of the data:

(a) if $S \leqslant \tau_1$, since $T \geqslant \tau_1$, then $D \geqslant \tau_1 - S$ (the likelihood contribution is $f(d|s \leqslant \tau_1) = f(d)/\{1 - F(\tau_1 - s)\}$);
(b) if $\tau_1 < S \leqslant \tau_2$, then the observation is unrestricted (since no delay is greater than 730 days) and the likelihood contribution is $f(d|\tau_1 < s \leqslant \tau_2) = f(d)$;

Table 3. Summary statistics for delay distributions (days)†

Organism name	*n*	*Minimum (days)*	Q_1 *(days)*	*Median (days)*	*Mean (days)*	Q_3 *(days)*	*Maximum (days)*
Acinetobacter baumanii	4022	1	8	14	41.32	31	721
Campylobacter jejuni	1977	2	7	14	25.81	22	726
Chlamydia sp	30758	2	9	14	34.01	24	728
Giardia lamblia	27166	0	6	9	25.43	18	729
Norovirus	56369	0	7	11	20.45	18	730
Salmonella abony	105	7	17	22	30.56	33	127
Salmonella braenderup	846	0	15	19	24.11	27	342
Salmonella brandenburg	176	8	18	25	50.0	50	631
Salmonella enteritidis PT21	3078	0	7	11	16.76	15	645
Salmonella infantis	1146	0	15	22	31.7	35	573
Salmonella senftenberg	478	4	13	18	24.15	26	420
Salmonella typhimurium DT104	2362	0	8	12	21.4	16	645

†n is the total number of isolates analysed, Q_1 is the 25% quantile and Q_3 the 75% quantile.

(c) if $\tau_2 < S \leqslant \tau_3$, since $T \leqslant \tau_3$, then $D \leqslant \tau_3 - S$ and the likelihood contribution is $f(d|\tau_2 < s \leqslant \tau_3) = f(d)/F(\tau_3 - s)$.

Hence, the log-likelihood given data s_i, t_i (with $d_i = t_i - s_i$), $i = 1, \ldots, n$, is

$$l = \sum_{i=1}^{n} \log \left\{ \frac{f(d_i)}{1 - F(\tau_1 - s_i)} J(s_i \leqslant \tau_1) + f(d_i)\, J(\tau_1 < s_i \leqslant \tau_2) + \frac{f(d_i)}{F(\tau_2 - s_i)} J(\tau_2 < s_i \leqslant \tau_3) \right\}, \qquad (2)$$

where J is the indicator function taking the value 1 if its argument is true and the value 0 otherwise.

The model will be parameterized in terms of the hazard $\lambda(d)$. Now

$$\lambda(d) = f(d)/S(d)$$

where the survivor function $S(d) = 1 - F(d)$ is

$$S(d) = \exp \left\{ - \int_0^d \lambda(u)\, \mathrm{d}u \right\}.$$

Thus the likelihood (2) can be rewritten in terms of the hazard function.

3.2. A semiparametric regression model for the hazard

We model the baseline hazard flexibly as a linear combination of M-splines; an M-spline of order o is a non-negative function made up of degree $o - 1$ polynomials connected at knots. Thus,

$$\lambda(d) = M(d) = \sum_{j=1}^{m} \alpha_j^2 M_j(d),$$

where the α_js are constants to be estimated, the squares ensuring that $M(d) \geqslant 0$. The $M_j(d)$, $j = 1, 2, \ldots, m$, are M-spline basis functions of order 4 and $m = o + k - 2$, where o is the order of the M-spline and k is the number of knots considered in the interval $[a, b]$ where $a = \min\{d_i, i = 1, \ldots, n\}$ and $b = \max\{d_i, i = 1, \ldots, n\}$ (including knots at a and b). M-splines and the implementation of the method have been described by Joly *et al.* (1998) and Ghebremichael-Weldeselassie *et al.* (2012), where further details may be sought.

In this paper, we take $k = 15$ (and hence $m = 17$). We found that using larger values of k only increases the computational complexity with no real improvement to the model. We constrained the model so that $\lambda(b) = 0$ by setting $\alpha_m = 0$ (Ramsay, 1988), consistent with observations with delay greater than 730 days being discarded; as it turned out, a lower discard time limit would have been appropriate for several organisms.

The advantage of using M-splines is that their integrals are I-splines, denoted $I(d)$. Thus the density may be written very simply as

$$f(d) = M(d)\,\exp\{-I(d)\}$$

and the CDF as

$$F(d) = 1 - \exp\{-I(d)\}.$$

We now extend the model to include a linear regression on the log-hazard scale. Let X be a vector of p covariates and θ be the corresponding vector of regression parameters. Let x_i be the covariate vector for observation i. Define

$$\begin{aligned}\lambda(d_i; x_i) &= \exp(\theta^{\mathrm{T}} x_i)\,\lambda(d_i)\\ &= \exp(\theta^{\mathrm{T}} x_i) \sum_{j=1}^{m} \alpha_j^2 M_j(d_i).\end{aligned}$$

Thus, the regression model is additive on the log-scale, the regression parameters being interpretable as log-relative-hazards.

The log-likelihood (2) is readily extended with $f(d_i; x_i)$ and $F(d_i; x_i)$ in place of $f(d_i)$ and $F(d_i)$ respectively. Let $l(\alpha, \theta; d, x)$ denote the log-likelihood of the semiparametric regression model, where α parameterizes the baseline hazard and θ the regression. Thus,

$$\begin{aligned}l(\alpha, \theta; d, x) = \sum_{i=1}^{n} \Bigg\{ &\left\{ l_i + \exp(\theta^{\mathrm{T}} x_i) \sum_{j=1}^{m} \alpha_j^2 I_j(\tau_1 - s_i) \right\} J(s_i \leqslant \tau_1) + l_i\, J(\tau_1 < s_i \leqslant \tau_2)\\ &+ \left(l_i - \log\left[1 - \exp\left\{ -\exp(\theta^{\mathrm{T}} x_i) \sum_{j=1}^{m} \alpha_j^2 I_j(\tau_2 - s_i) \right\} \right] \right) J(\tau_2 < s_i \leqslant \tau_3) \Bigg\}, \qquad (3)\end{aligned}$$

where

$$l_i = \theta^{\mathrm{T}} x_i + \log\left\{ \sum_{j=1}^{m} \alpha_j^2 M_j(d_i) \right\} - \exp(\theta^{\mathrm{T}} x_i) \sum_{j=1}^{m} \alpha_j^2 I_j(d_i).$$

The parameters α and θ are obtained by maximizing the penalized log-likelihood, which involves a smoothing parameter ϕ that controls the curviness of $M(d)$:

$$\mathrm{pl}(\alpha, \theta; d, x) = l(\alpha, \theta; d, x) - \phi \int \left\{ \sum_{j=1}^{m} \alpha_j^2 M_j''(u) \right\}^2 \mathrm{d}u.$$

The smoothing parameter ϕ is obtained by maximizing an approximate cross-validation criterion (O'Sullivan, 1988). As proposed by Joly *et al.* (1998) and also used by Ghebremichael-Weldeselassie *et al.* (2012), ϕ is chosen in the absence of covariates (and thus with θ set to 0). We found that the results are not very sensitive to the choice of ϕ.

We also undertook analyses of the hazard on subsets of the data, segmented in two groups: short delays ($D \leqslant 60$ days) and longer delays ($D > 60$ days). For the short delays group we estimated the hazard non-parametrically as previously described, using new values τ_1 and τ_2 in

the likelihood corresponding to a maximum delay of 60 days rather than 730 days as previously used. For the longer delays we fitted an exponential model, conditional on $D > 60$, applying the same likelihood as described previously, restricted to values $d_i > 60$, and using the conditional exponential density $f(d|D > 60) = \lambda \exp\{-\lambda(d - 60)\}$ and CDF $F(t|D > 60) = 1 - \exp\{-\lambda(d - 60)\}$. Regression models were not fitted to these subsets of the data.

4. Data analysis

We first fitted the model without any covariates to the data for the 12 pathogens listed in Table 1. We also did a combined analysis of the five relatively infrequent salmonellas (*Salmonella abony*, *Salmonella braenderup*, *Salmonella brandenburg*, *Salmonella infantis* and *Salmonella senftenberg*), since these organisms have similar delay distributions. Figs 1 and 2 show the histogram of the delays and the estimated hazards. For all data sets, the histograms reveal that the delay distributions have very long upper tails, with most delays being less than 2 months. Generally, the hazards are highly peaked at low values and then drop dramatically to fluctuate around a low positive value. There are two exceptions to this general pattern: *Chlamydia* sp, for which the drop in the hazard at long delays is followed by a very large peak at about 400–500 days, and the grouped uncommon salmonellas, for which the hazard peaks at 100 days. Norovirus also exhibits a lesser secondary peak in the hazard at a delay of 150–200 days. These peaks are due to irregularities in the upper tail of the distribution, which are discussed below, possibly resulting from organisms with delays over a certain value being discarded.

A roughly constant hazard at long delays suggests that, beyond a certain point, reporting is essentially memoryless (and hence consistent with an exponential distribution). This accords with the interpretation that long delays are either the result of random coding errors or relate to specimens that have been lost, set aside or overlooked and are reported when they happen to turn up, or when cleared out. The highly peaked hazard at low delays, in contrast, reflects the laboratory identification process: there are relatively few very short delays, at which point the hazard is low, thereafter increasing to a peak before dropping again.

To look in more detail at these two very different features of the delay distributions, we repeated the analyses after segmenting the data in two groups: short delays ($D \leqslant 60$ days) and longer delays ($D > 60$ days). We chose a cut-off of 60 days because reports with delays beyond this are of little use in outbreak detection; for eight of the 12 organisms, the value of 60 days lies beyond the 90th percentile of the delay distribution. Figs 3 and 4 show the histograms of the delays in the two groups, along with the fitted non-parametric and exponential densities.

On this more detailed scale, it is clear that the empirical delay distribution under 60 days can be markedly multimodal (which is a feature generally smoothed over by the spline model, notably for the uncommon salmonellas; see Fig. 4(d), with modes typically at weekly intervals. This is particularly apparent for *Campylobacter jejuni* and the salmonellas. The weekly modes may reflect the practice of some laboratories, which may work to a weekly reporting cycle; further enquiries suggest that laboratory reporting practice is extremely variable and idiosyncratic. The modes may also reflect approximate coding of dates. The upper tails, in contrast, often reveal irregularities. Most notable are reporting cliffs, i.e. sudden and permanent drops in frequency. This is apparent for *Chlamydia* sp, for norovirus and for some of the less common salmonellas, e.g. *Salmonella senftenberg*. These discontinuities in the delay distribution cause fluctuations in the hazards in the upper tail, which are visible for certain organisms in Figs 1 and 2. It appears that some laboratories do not report organisms with delays beyond a certain value: the hazard should be zero from that value onwards. Under this interpretation, the fluctuations in

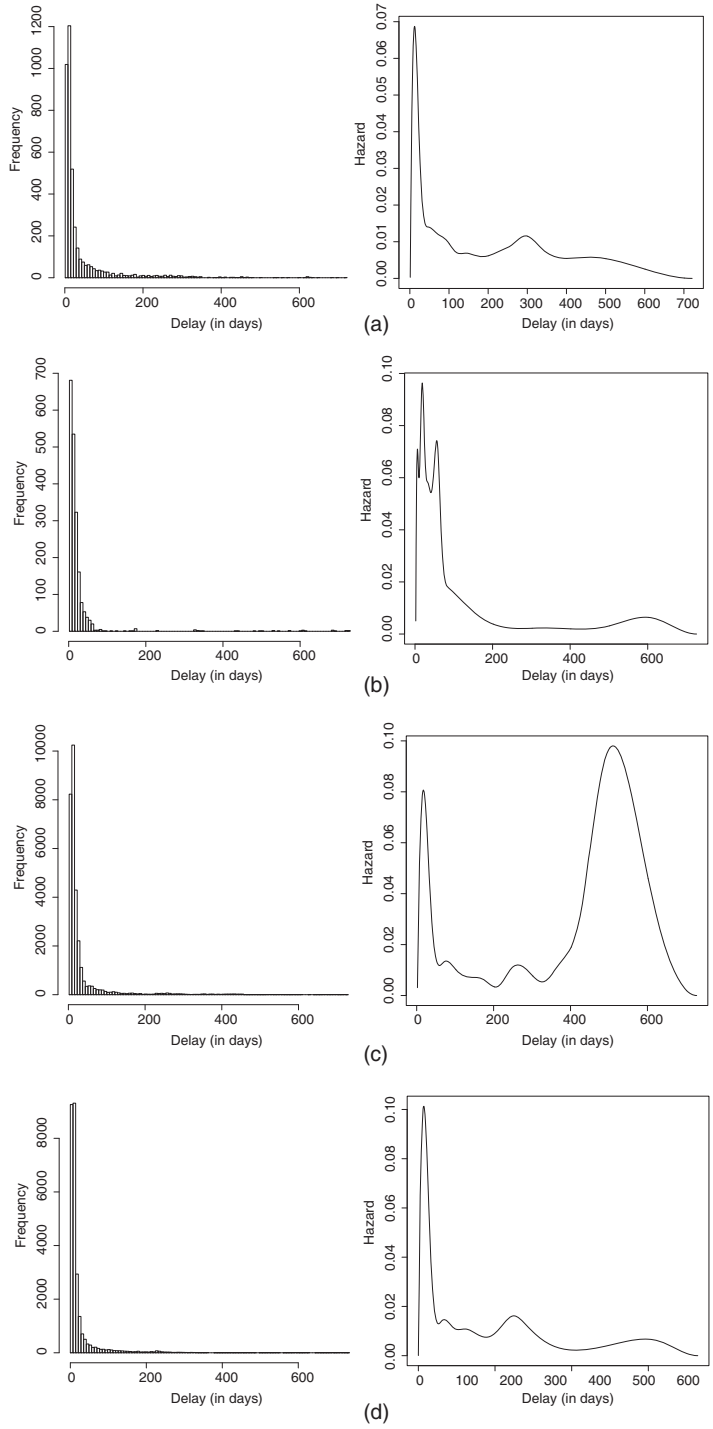

Fig. 1. Histogram and estimated smooth hazard of the delay distribution for four organisms: (a) *Acinetobacter baumanii*; (b) *Campylobacter jejuni*; (c) *Chlamydia* sp; (d) *Giardia lamblia*

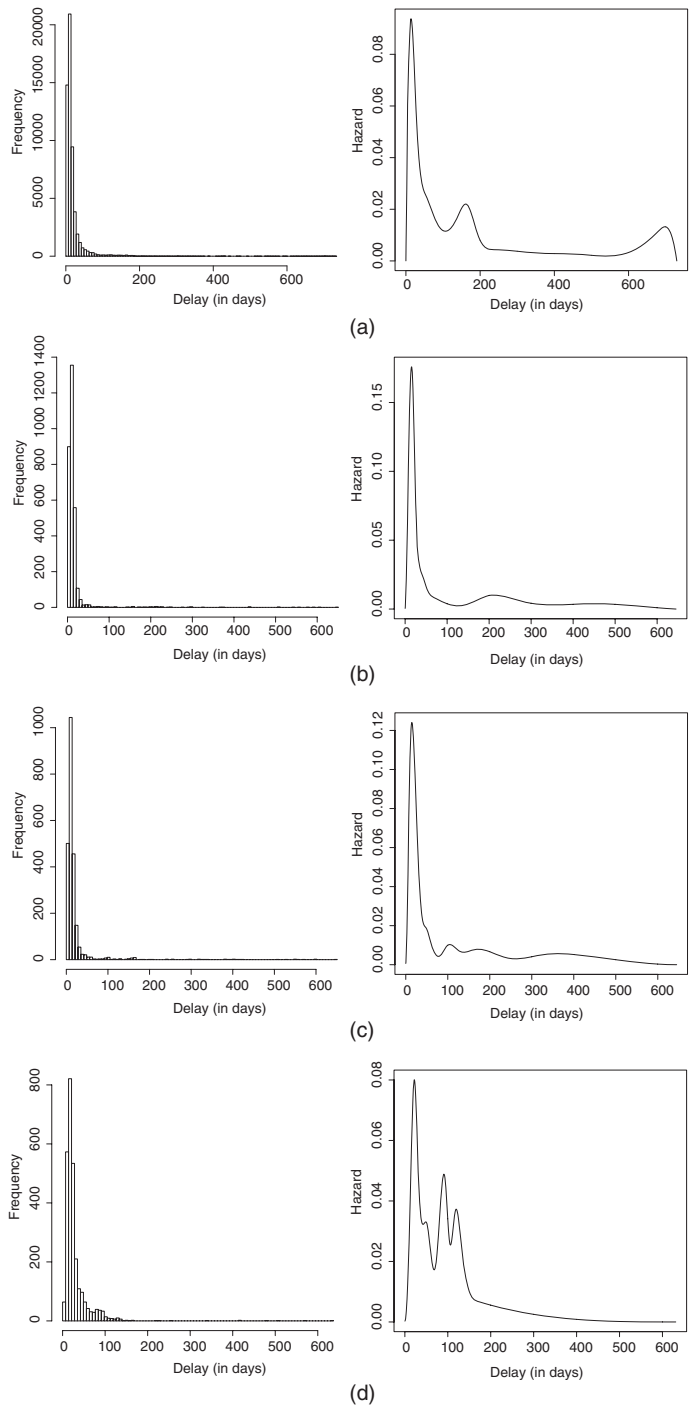

Fig. 2. Histogram and estimated smooth hazard of the delay distribution for three organisms and the five uncommon salmonellas combined: (a) norovirus; (b) *Salmonella enteritidis* PT21; (c) *Salmonella typhimurium* DT104; (d) uncommon salmonellas

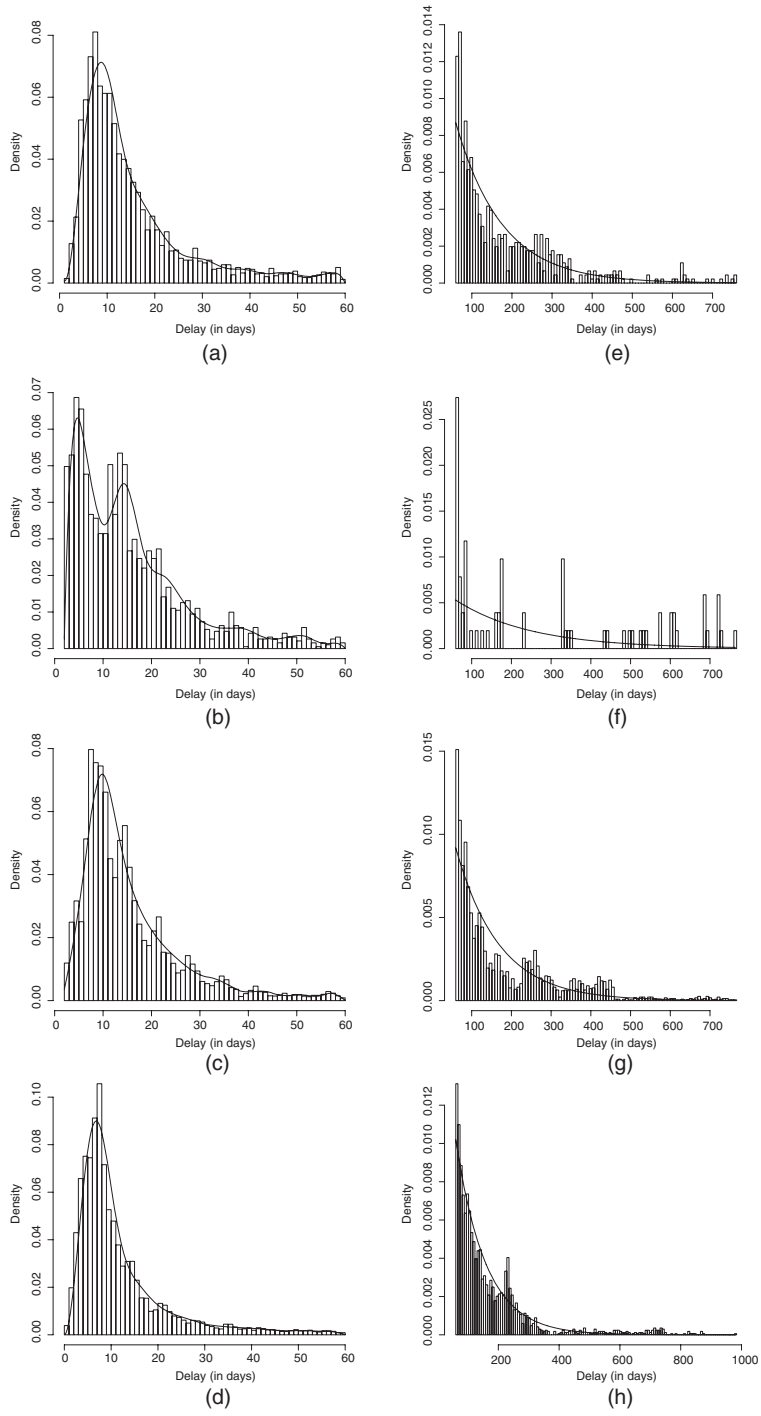

Fig. 3. Histogram and fitted density for four organisms, for (a)–(d) delays of 60 days or less, with the density obtained from the smooth hazard, and (e)–(h) delays greater than 60 days, with the exponential density: (a), (e) *Acinetobacter baumanii*; (b), (f) *Campylobacter jejuni*; (c), (g) *Chlamydia* sp; (d), (h) *Giardia lamblia*

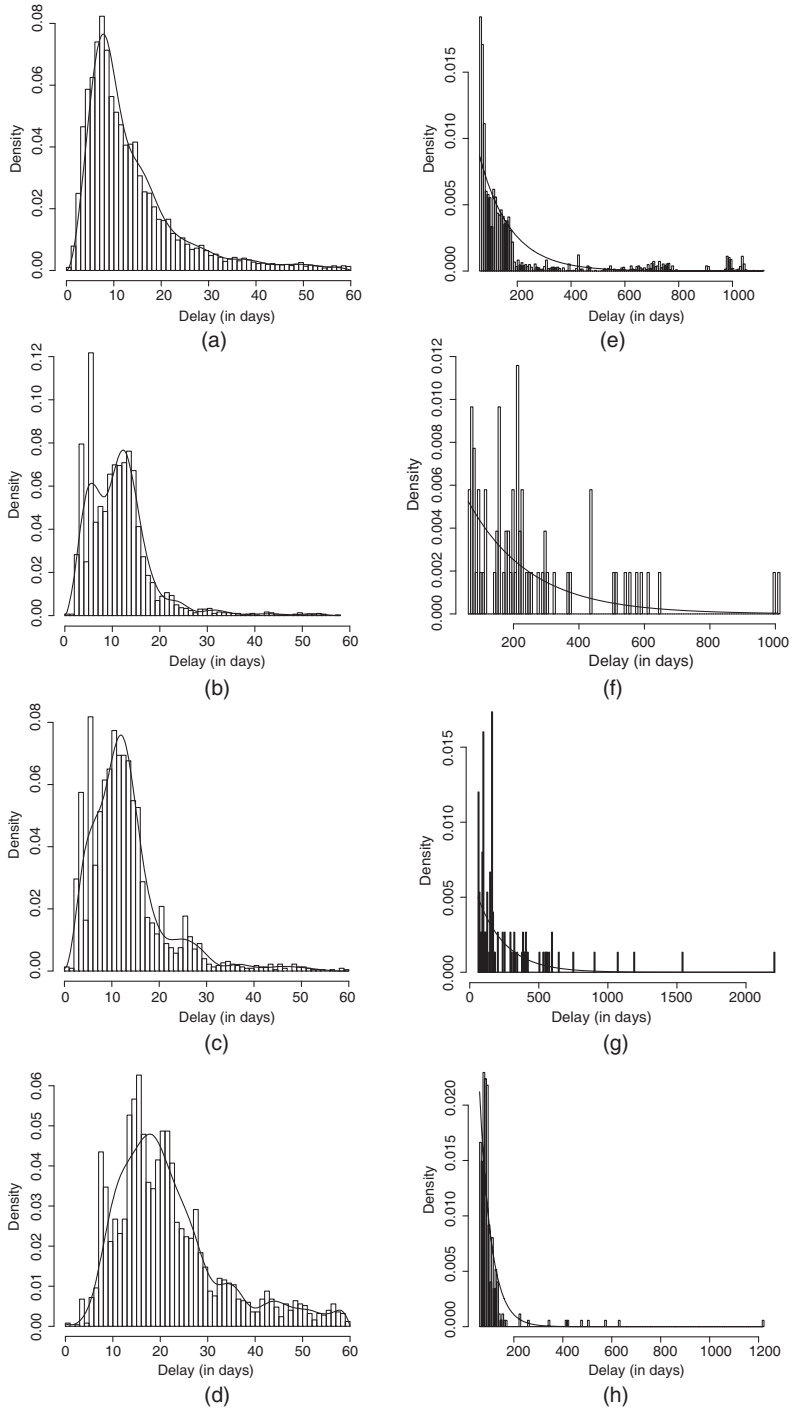

Fig. 4. Histogram and fitted density for three organisms, and the five uncommon salmonellas combined for (a)–(d) delays of 60 days or less, with the density obtained from the smooth hazard, and (e)–(h) delays greater than 60 days, with the exponential density: (a), (e) norovirus; (b), (f) *Salmonella enteritidis* PT21; (c), (g) *Salmonella typhimurium* DT104; (d), (h) uncommon salmonellas

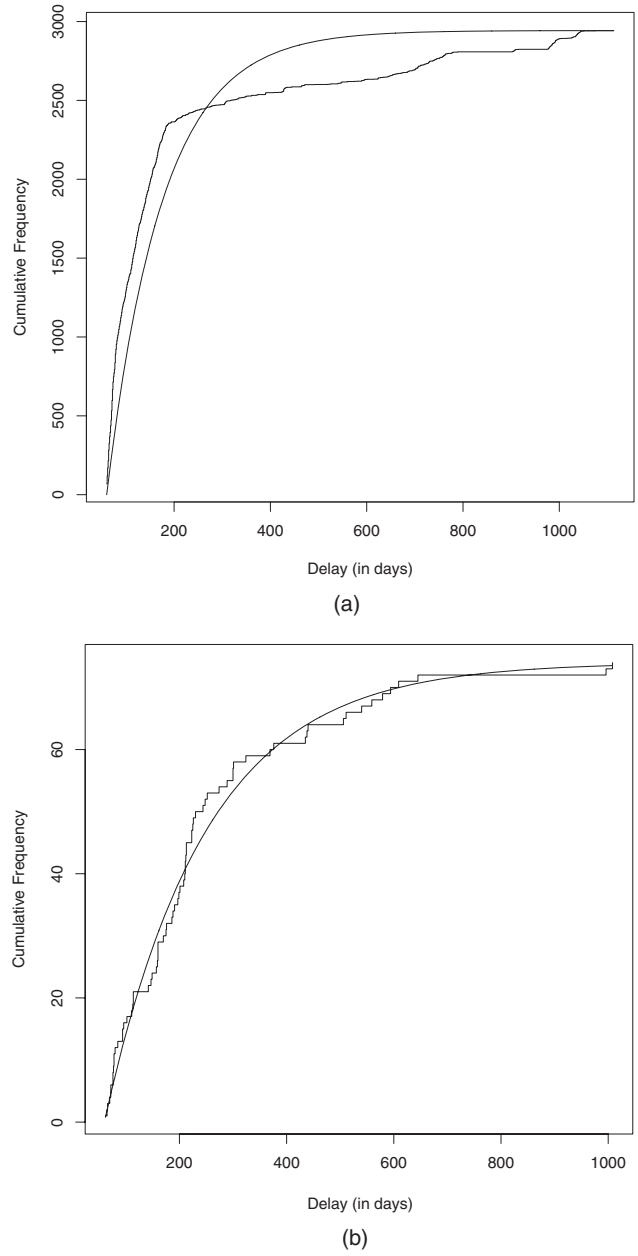

Fig. 5. Observed and fitted exponential cumulative frequency of delays greater than 60 days for two organisms: (a) norovirus; (b) *Salmonella enteritidis* PT21

the hazard at these points is spurious. In view of these irregularities, the exponential model does not generally provide a statistically acceptable fit to the upper tail, though it often adequately describes the overall tendency. Plots of the empirical and expected (under the exponential model) cumulative frequency of delays above 60 days are shown in Fig. 5, for norovirus where the exponential model is deficient, and *Salmonella enteritidis* PT21, where it is adequate.

Returning to the model for all delays from 0 to 730 days, we next investigated the effect of

covariates on the hazard by using the semiparametric regression model. We looked for linear and quadratic effects of calendar time, using the covariates x_1 and $x_2 = x_1^2$, with

$$x_{1i} = (s_i - m)/1000$$

where s_i is the date of specimen collection for observation i and m is the mean of the s_i for the organism under consideration. We looked at the effect of season with covariates x_3, x_4 and x_5 representing indicators for summer (s_i in June–August), autumn (s_i in September–November) and winter (s_i in December–February) respectively, the reference season being spring. Finally, we looked at the effect of recent throughput for that organism, using the covariate x_6 defined as

$$x_{6i} = \frac{1}{100} \times \text{ total number of specimens collected in the 7 days before } s_i.$$

The log-likelihood ratios for fitting these covariates one at a time and jointly is shown in Table 4. When each covariate is assessed separately at the 5% level of significance, there was no evidence of any trend for *Salmonella braenderup* or *Salmonella senftenberg*; for *Salmonella abony*, *Salmonella brandenburg* and *Salmonella infantis* there was evidence of a linear trend (on the log-scale) but not of a quadratic trend. For *Acinetobacter baumanii*, *Salmonella abony* and *Salmonella brandenburg* there was no significant season effect. For *Chlamydia* sp, *Salmonella abony*, *Salmonella infantis* and *Salmonella typhimurium* DT104 there was little evidence that the number of specimens of the same organism collected in the previous week had any bearing on results. For *Chlamydia* sp, the full model did not converge so we replaced the covariate x_{6i} by an indicator taking the value 1 when the specimen count in the previous week was above its median value. We also analysed the five less common salmonellas combined. Table 5 shows the parameter estimates that were obtained when all covariates (including the quadratic trend) are included.

To assess the effect on the hazard of the previous week's specimen count, we calculated $\delta = \exp(\theta \, \mathrm{IQR}/100)$ where IQR is the interquartile range of the weekly specimen count. These

Table 4. Likelihood ratio test statistics relative to the null model†

Organism name	*Results for the following models and numbers of parameters:*				
	Linear trend, 1	*Quadratic trend, 2*	*Season effect, 3*	*Previous weekly count, 1*	*All variables, 6*
Acinetobacter baumanii	45.3	113.9	*1.3*	7.4	138.5
Campylobacter jejuni	50.3	56.5	41.3	32.2	175.6
Chlamydia sp	81.4	1034.2	285.6	*2.8*‡	1695.2
Giardia lamblia	3582.4	3604.8	165.2	141.7	3971.6
Norovirus	7307.2	7528.0	44.4	2925.6	8286.0
Salmonella abony	6.6	6.9	*2.7*	*3.0*	14.0
Salmonella braenderup	*2.3*	*2.3*	44.6	5.4	59.3
Salmonella brandenburg	25.5	26.2	26.4	*4.0*	61.1
Salmonella enteritidis PT21	24.0	103.9	35.5	195.0	236.9
Salmonella infantis	117.9	118.3	59.9	9.3	166.6
Salmonella senftenberg	*0.2*	*4.6*	13.9	21.1	23.2
Salmonella typhimurium DT104	*0.2*	47.9	38.5	*0.2*	96.8
5 infrequent salmonellas	111.3	113.2	153.4	27.4	284.0

†Values with $p > 0.05$ are in italics.
‡Dichotomous variable (see the text).

Table 5. Parameter estimates (with standard errors in parentheses below them) and value of δ for the joint models

Organism name	*Results for the following covariates:*						
	Trend		*Season*			*Recent incidence*	
	Linear, β_1	*Quadratic,* β_2	*Summer,* γ_1	*Autumn,* γ_2	*Winter,* γ_3	*Count,* θ	*(See text),* δ
Acinetobacter baumanii	0.222	0.221	−0.012	−0.029	−0.047	2.01	1.13
	(0.021)	(0.027)	(0.046)	(0.046)	(0.048)	(0.433)	
Campylobacter jejuni	0.275	0.101	−0.125	−0.701	−0.154	1.25	1.08
	(0.028)	(0.029)	(0.060)	(0.074)	(0.070)	(0.23)	
Chlamydia sp	0.268	−0.368	−0.085	−0.022	0.241	24.3†	1.28
	(0.010)	(0.011)	(0.016)	(0.017)	(0.017)	(1.51)	
Giardia lamblia	0.404	0.020	−0.075	−0.001	0.012	0.738	1.19
	(0.007)	(0.009)	(0.014)	(0.014)	(0.008)	(0.015)	
Norovirus	0.505	0.142	−0.016	−0.025	−0.122	0.066	1.16
	(0.007)	(0.007)	(0.017)	(0.014)	(0.011)	(0.002)	
Salmonella abony	0.339	−0.055	−0.462	−0.690	−0.231	44.9	1.00
	(0.129)	(0.158)	(0.264)	(0.262)	(0.343)	(24.1)	
Salmonella braenderup	0.086	−0.017	−0.332	−0.610	0.026	4.95	1.16
	(0.043)	(0.056)	(0.104)	(0.099)	(0.119)	(1.62)	
Salmonella brandenburg	0.654	−0.237	1.138	0.581	2.031	14.4	1.15
	(0.118)	(0.143)	(0.465)	(0.443)	(0.511)	(10.3)	
Salmonella enteritidis PT21	0.011	−0.238	0.015	0.059	0.057	2.08	1.24
	(0.033)	(0.042)	(0.055)	(0.054)	(0.046)	(0.217)	
Salmonella infantis	0.402	0.010	−0.184	−0.465	0.081	1.44	1.21
	(0.041)	(0.052)	(0.093)	(0.091)	(0.105)	(1.48)	
Salmonella senftenberg	−0.111	−0.188	−0.263	−0.496	−0.073	6.22	1.04
	(0.076)	(0.083)	(0.116)	(0.105)	(0.139)	(1.77)	
Salmonella typhimurium DT104	0.131	−0.301	−0.018	0.323	0.240	0.507	1.04
	(0.036)	(0.043)	(0.067)	(0.069)	(0.073)	(0.161)	
5 infrequent salmonellas	0.217	−0.060	−0.349	−0.597	0.038	2.52	1.16
	(0.027)	(0.032)	(0.063)	(0.061)	(0.065)	(0.462)	

†Dichotomous variable (see the text).

values are in Table 5. Thus, for each organism, δ is the hazard at the 75th percentile of x_{6i}, divided by the hazard at the 25th percentile of x_{6i} (for *Chlamydia* sp, $\delta = \exp(\theta/100)$). For example, for *Acinetobacter baumanii*, the lower and upper quartiles of the weekly counts are 7 and 13 respectively, so $\delta = \exp\{2.01(13-7)/100\} \simeq 1.13$. This δ provides a more interpretable measure than θ and permits comparisons between organisms of different frequencies. The hazard ratios δ all lie between 1 and 1.3, indicating an increase in the hazard and hence a reduction in the delay. However, the effect of recent reports is not as marked as that of calendar time. Note that this analysis considers workload related to only the same organism: an alternative would have been to consider total workload irrespective of organism type.

Fig. 6 shows the trends over time for all organisms (reparameterized to share a common reference point); the five uncommon salmonellas were grouped. For most organisms, notably norovirus, the hazard has increased in recent years, corresponding to a reduction in reporting delays, though for *Chlamydia* sp, *Salmonella typhimurium* DT104 and *Salmonella enteritidis* PT21 there has been a slight reduction in the hazard, and hence a lengthening of the delay. Changes in reporting delay distributions over time may be affected by the changing efficiency of

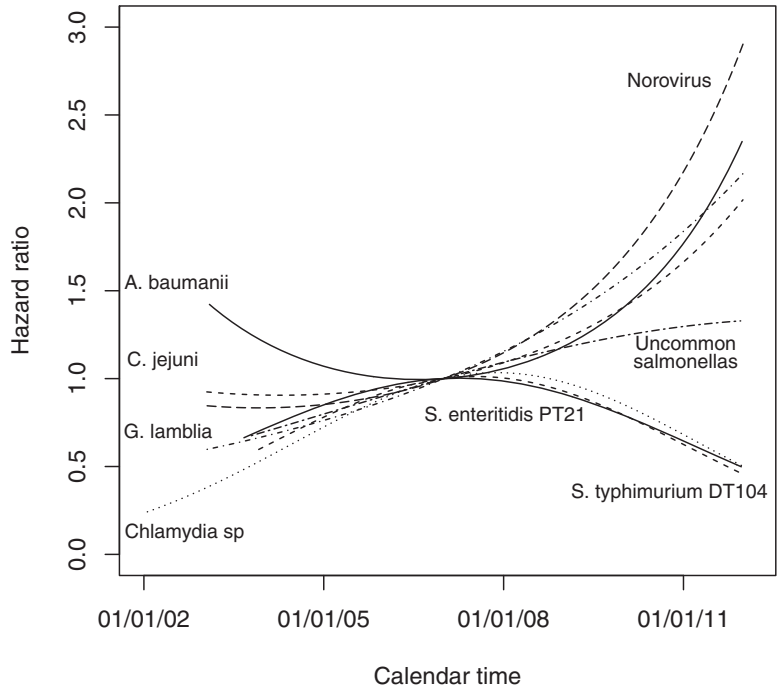

Fig. 6. Quadratic trend for seven organisms and the uncommon salmonellas combined

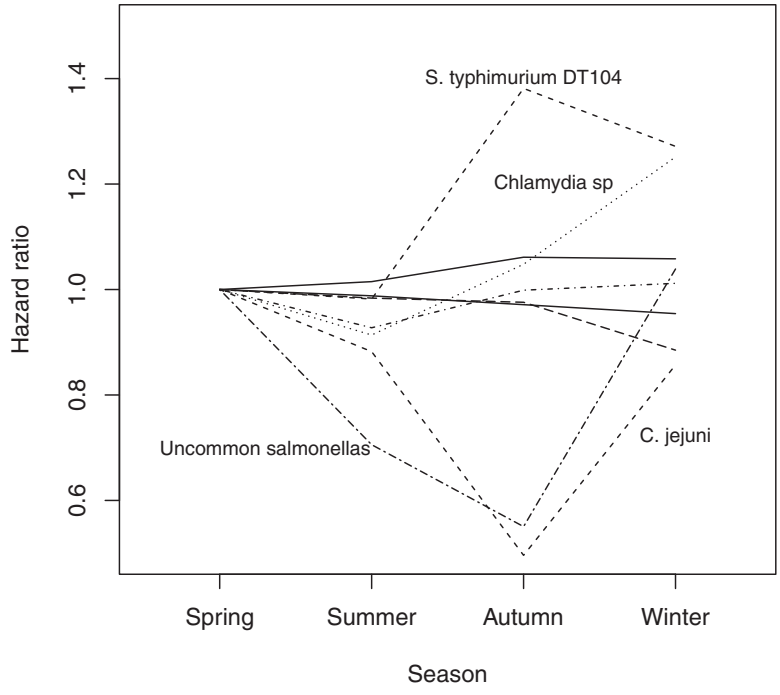

Fig. 7. Seasonal effect for seven organisms and the uncommon salmonellas combined

testing procedures, and by public health priorities. The increasing hazard (and hence decreasing delays) for norovirus may be affected by the growing public profile that has been accorded to winter vomiting disease.

Fig. 7 shows the seasonal effect for all organisms (with the five infrequent salmonellas grouped). It is clear that, although there are some seasonal effects, these are much less marked than the effects of calendar time.

We examined the proportional hazards assumption by plotting the empirical cumulative hazards $\log[-\log\{\tilde{S}(d|x)\}]$ against the delay d for specimen dates between 2002 and 2009 (to avoid censoring), where $\tilde{S}$ denotes the empirical survivor function and x defines a subgroup of specimen dates (the plots are not shown). When the estimated hazard ratios were substantially different from 1, the proportional hazard assumption was generally found to be plausible, as for example in the case of the time trends in norovirus. For relative hazards that were close to 1, the corresponding cumulative hazards tended to be overlaid, the noisiness of the data precluding any assessment of the proportionality assumption.

5. Final remarks

We have analysed reporting delays for a selection of organisms reported to the LabBase surveillance database. Two major features emerge from this analysis. First, reporting delays can be extremely long. Although we excluded data with reporting delays over 2 years as these were very uncommon indeed, long delays remain a problem. The shape of the hazard suggests that, after a certain point, reporting becomes essentially random, with some laboratories stopping entirely to report organisms with delays beyond a certain time. Most delays, however, lie within a relatively short window, determined by the organism identification procedure but typically under 2 months. Delays in this range are relevant to outbreak detection: data with longer delays are essentially irrelevant for this purpose. The second feature to emerge is that temporal effects do influence the delay distribution. Most important among these are long-term trends over calendar time, which correspond to a gradual reduction or lengthening of the delays. The practical consequence of this observation is that the delay distribution cannot be assumed to be stationary over long time spans.

The modelling framework that we have used is based on a semiparametric regression model for the hazard. This enabled us to visualize the hazard in continuous time, and to study the effect of covariates in a natural fashion, with a simple relative hazard interpretation. The major disadvantage of the approach that we have taken is that it is more cumbersome than the generalized linear model method of Brookmeyer and Liao (1990). Estimates of the hazard at long delays are sensitive to irregularities in the data, notably those resulting from discontinuities in reporting. Such sensitivities in the hazard, though often spurious, can help to focus attention on important aspects of the data.

Our emphasis throughout this paper has been to describe and characterize the delay distribution in its entirety. In particular, we have included long delays, which are often dropped from the analysis of surveillance data. Such an omission is arguably perverse, since a better understanding of these long delays may perhaps contribute to improving laboratory reporting practice, and hence to reducing both the extremes and the means of the delay distributions. For detection of outbreaks, reporting delays constitute a nuisance factor that may delay the identification of an outbreak, or blur its magnitude. Isolates with long delays are of little use in outbreak detection, since, by the time that they are reported, opportunities for control of the outbreak may have passed, though of course this in turn depends on the duration of the outbreak. More emphasis should therefore be placed on eliminating extreme delays.

Acknowledgements

This research was supported by a grant from the UK Medical Research Council. CPF was supported by a Royal Society Wolfson Research Merit Award.

References

Bacchetti, P. (1996) Reporting delays of deaths with AIDS in the United States. *J. Acq. Immun. Defic. Synd. Hum. Retrovir.*, **13**, 363–367.

Brookmeyer, R. and Damiano, A. (1989) Statistical methods for short-term projections of AIDS incidence. *Statist. Med.*, **8**, 23–34.

Brookmeyer, R. and Gail, M. (1988) A method for obtaining short-term projections and lower bounds on the size of the Aids epidemic. *J. Am. Statist. Ass.*, **83**, 301–308.

Brookmeyer, R. and Liao, J. (1990) The analysis of delays in disease reporting: methods and results for the acquired immunodeficiency syndrome. *Am. J. Epidem.*, **122**, 355–365.

Cui, J. and Kaldor, J. (1998) Changing pattern of delays in reporting AIDS diagnoses in Australia. *Aust. New Zeal. J. Publ. Hlth*, **22**, 432–435.

Deuffic, S. and Costagliola, D. (1999) Is the AIDS incubation time changing?: a back-calculation approach. *Statist. Med.*, **18**, 1031–1047.

Enki, D. G., Noufaily, A., Garthwaite, P. H., Andrews, N. J., Charlett, A. and Farrington, C. P. (2013) Automated biosurveillance data from England and Wales, 1991 - 2011. *Emergng Infect. Dis.*, **19**, 35–42.

Farrington, C. P., Andrews, N. J., Beale, A. D. and Catchpole, M. A. (1996) A statistical algorithm for the early detection of outbreaks of infectious disease. *J. R. Statist. Soc.* A, **159**, 547–563.

Gebhardt, M. D., Neuenschwander, B. E. and Zwahlen, M. (1998) Adjusting AIDS incidence for non-stationary reporting delays: a necessity for country comparisons. *Eur. J. Epidem.*, **14**, 595–603.

Ghebremichael-Weldeselassie, Y., Whitaker, H. J. and Farrington, C. P. (2012) Self-controlled case series method with smooth age effect. *Technical Report 12/07*. The Open University, Milton Keynes. (Available from `http://statistics.open.ac.uk/`.)

Harris, J. E. (1990) Reporting delays and the incidence of AIDS. *J. Am. Statist. Ass.*, **85**, 915–924.

Joly, P., Commenges, D. and Letenneur, L. (1998) A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia. *Biometrics*, **54**, 185–194.

Jones, G., Le Hello, S., Jourdan-da Silva, N., Vaillant, V., de Valk, H., Weill, F.-X. and Le Strat, Y. (2014) The French human salmonella surveillance system: evaluation of timeliness of laboratory reporting and factors associated with delays, 2007 to 2011. *Eurosurveillance*, **14**, no. 1.

Kalbfleisch, J. D. and Lawless, J. F. (1989) Inference based on retrospective ascertainment: an analysis of data on transfusion related to AIDS. *J. Am. Statist. Ass.*, **84**, 360–372.

Lawless, J. F. (1994) Adjustments for reporting delays and the prediction of occurred but not reported events. *Can. J. Statist.*, **22**, 15–31.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Midthune, D. N., Fay, M. P., Clegg, L. X. and Feuer, E. J. (2005) Modelling reporting delays and reporting corrections in Cancer registry data. *J. Am. Statist. Ass.*, **100**, 61–70.

Noufaily, A., Enki, D. G., Farrington, P., Garthwaite, P., Andrews, N. and Charlett, A. (2013) An improved algorithm for outbreak detection in multiple surveillance systems. *Statist. Med.*, **32**, 1206–1222.

O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M. and Mollison, D. (2000) Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Appl. Statist.*, **49**, 517–542.

O'Sullivan, F. (1988) Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Scient. Statist. Comput.*, **9**, 363–379.

Ramsay, J. O. (1988) Monotone regression splines in action. *Statist. Sci.*, **3**, 425–461.

Sartwell, P. E. (1966) The incubation period and the dynamics of infectious disease. *Am. J. Epidem.*, **83**, 204–216.

Tabnak, F., Müller, H.-G., Wang, J.-L., Chiou, J.-M. and Sun, R. K. P. (2000) A change-point model for reporting delays under change of AIDS case definition. *Eur. J. Epidem.*, **16**, 1135–1141.

Zeger, S. L., See, L. and Diggle, P. J. (1989) Statistical methods for monitoring the AIDS epidemic. *Statist. Med.*, **8**, 3–21.