Bayesian Nowcasting during the STEC O104:H4 Outbreak in Germany, 2011

Michael Höhle[1,2,*] and Matthias an der Heiden[2]

[1]Department of Mathematics, Stockholm University, Sweden
[2]Department for Infectious Disease Epidemiology, Robert Koch Institute, Berlin, Germany

This is an author-created preprint of the article which has been accepted for publication in Biometrics as DOI 10.1111/biom.12194. The definite version is available from onlinelibrary.wiley.com. The on-line appendices as well as the definite version can also be found directly as http://dx.doi.org/10.1111/biom.12194.

Abstract

A Bayesian approach to the prediction of occurred-but-not-yet-reported events is developed for application in real-time public health surveillance. The motivation was the prediction of the daily number of hospitalizations for the haemolytic-uremic syndrome during the large May-July 2011 outbreak of Shiga toxin-producing *Escherichia coli* (STEC) O104:H4 in Germany. Our novel Bayesian approach addresses the count data nature of the problem using negative binomial sampling and shows that right-truncation of the reporting delay distribution under an assumption of time-homogeneity can be handled in a conjugate prior-posterior framework using the generalized Dirichlet distribution. Since, in retrospect, the true number of hospitalizations is available, proper scoring rules for count data are used to evaluate and compare the predictive quality of the procedures during the outbreak. The results show that it is important to take the count nature of the time series into account and that changes in the delay distribution occurred due to intervention measures. As a consequence, we extend the Bayesian analysis to a hierarchical model, which combines a discrete time survival regression model for the delay distribution with a penalized spline for the dynamics of the epidemic curve. Altogether, we conclude that in emerging and time-critical outbreaks, nowcasting approaches are a valuable tool to gain information about current trends.

Keywords: Infectious disease epidemiology; Reporting delay; Truncation; Real-time surveillance.

1 Introduction

During May–July 2011, Germany was confronted with a large outbreak of gastrointestinal disease caused by Shiga toxin-producing *Escherichia coli* (STEC) O104:H4 associated with sprouts consumption. A total of 2,987 cases of diarrhea without the haemolytic uremic syndrome (HUS) complication and 855 cases of HUS were attributable to the outbreak,

*Author of correspondence: Department of Mathematics, Stockholm University, Kräftriket, 106 91 Stockholm, Sweden, Email: hoehle@math.su.se

making this one of the largest STEC outbreaks ever reported (Frank et al., 2011; Buchholz et al., 2011). During the outbreak it was vital to have daily information on current epidemic trends in order to judge, if the outbreak was ongoing, assess the impact of control measures and perform capacity planning. However, such real-time tracking is complicated by the inherent delay in public health reporting systems between the occurrence of the event, e.g. time of symptom onset or hospitalization, and the time the report becomes available in the public health surveillance database. Following Donker et al. (2011) such delay-adjusting tracking procedures are called *nowcasts* in the public health setting.

We address the nowcasting task in the statistical framework of the occurred-but-not-reported-events problem (Lawless, 1994). Here, estimation of the delay distribution takes the inherent right-truncation of the data generating process into account. The problem originates from actuarial science, where it is known as claims reserving modelling (England and Verrall, 2002), but has also found application in a biostatistical context when analysing the AIDS/HIV epidemic (Brookmeyer and Damiano, 1989; Kalbfleisch and Lawless, 1989; Zeger et al., 1989). Applications can also be found in non-infectious disease modelling such as cancer registry data (Midthune et al., 2005) or mortality monitoring (Lin et al., 2008). Besides the already mentioned use during the 2009 H1N1 influenza pandemic in Donker et al. (2011), nowcasting procedures have also been used to assess excess mortality during heatwaves (Green et al., 2012) and – in a somewhat different form – to perform influenza-like illness surveillance (Nunes et al., 2013).

Figure 1 shows the daily number of HUS hospitalizations during the outbreak; 658 of the 827 (i.e. 79.6%) now confirmed HUS cases have information on the date of hospitalization available. Since hospitalization can be assumed for HUS cases, missingness merely reflects that the event occurrence was not recorded, e.g., because it is not a mandatory reporting field or, because it was not possible to ascertain with sufficient precision (for the 169 HUS cases without hospitalization date, 101 were recorded as having been hospitalized, 29 had missing information on this matter and only 39 were indicated as not having been hospitalized). Also shown in the figure is the number of hospitalizations available in the central German surveillance database SurvNet@RKI as of 2011-06-02 – the discrepancy between the two curves is due to reporting delay: the notification from the hospital goes to the local health department from which it is then forwarded to the Robert Koch Institute (RKI) via the state health department. The goal of nowcasting is to predict the true number of counts from the currently available counts. The corresponding animation of available reports as a function of time (in days) in the period 2011-05-01 to 2011-07-06 can be found as Web Animation 1.

The aim of our work is – in hindsight – to evaluate the quality of the nowcasting procedures on HUS hospitalizations during the outbreak. We chose the date of hospitalization for HUS as the target, since it constituted a more reliable and outbreak specific indicator compared to, e.g., onset of symptoms in all STEC cases. In the present work we propose a novel Bayesian hierarchical model for the nowcasting problem, which better reflects uncertainty due to both estimation of the delay distribution and the count data nature of the hospitalization data. A unique feature of our work is that, on the basis of the full epidemic curve, we are able to retrospectively perform a quantitative evaluation of different nowcast schemes based on proper scoring rules (Czado et al., 2009). This allows us to address both aspects of sharpness and calibration of the nowcasts and permits a ranking of the procedures.

Our work is structured as follows: Section 2 introduces nowcasting and its mathematical notation. Section 3 considers frequentist and Bayesian nowcasting when the delay distribution can be assumed as time-homogeneous, whereas Sect. 4 contains a joint Bayesian modelling of both time-varying delay distribution and epidemic curve. Proper scoring rules, which we use

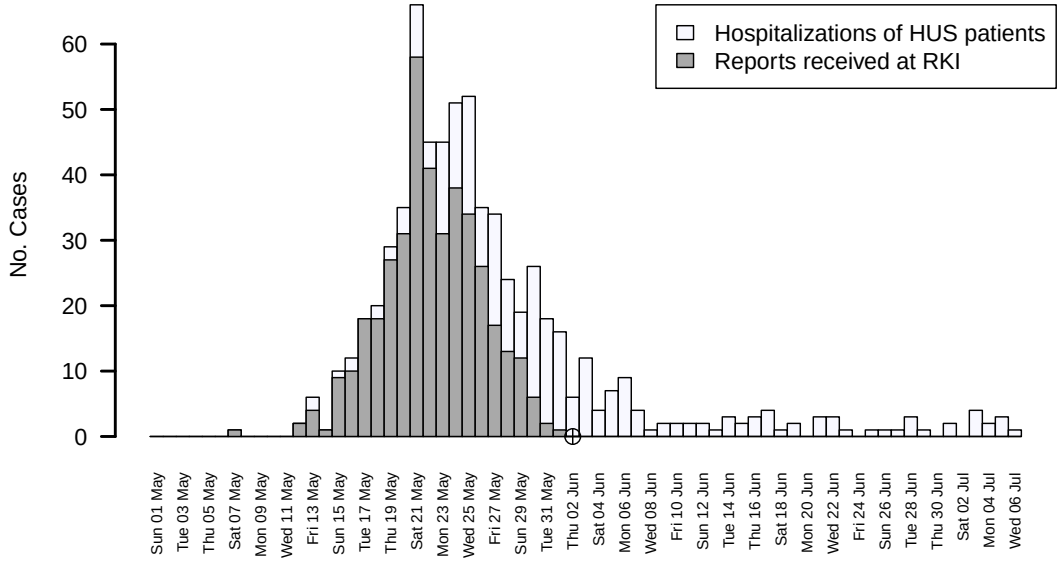

Figure 1: Daily number of hospitalizations due to HUS during the outbreak as available in retrospect. Also shown in darkgray are the number of available hospitalization reports at the RKI as of 2011-06-02 (indicated by the chrosshairs symbol).

to evaluate the predictive quality of the casts, are explained in Sect. 5. In Sect. 6 all methods are applied and evaluated on the STEC outbreak data. Finally, a discussion of the results and an outlook in Sect. 7 completes the work.

2 Nowcasting

We adapt the notation of Lawless (1994) to describe the nowcasting problem. Let $n_{t,d}$ be the number of cases which occur on day t and become available with a delay of d days, i.e. the case reports arrive on day $t+d$. Here, the indices span $t = 0, \ldots, T$, with T being the index of the current day, i.e. *now*, and $d = 0, \ldots, D$. It is typical to assume that delays can occur only up to a maximum of D days; e.g., because larger delays can not be estimated reliably or because cases with a longer delay provide information about times too far in the past to be relevant. Hence, the last category corresponding to D days often covers the case 'D days or more'. We shall return to this issue in Section 6. While nowcasting can be performed repeatedly for different "nows" as described in Sections 5 and 6, the current description focuses on a single "now" T. Figure 2 illustrates the *reporting triangle* of $n_{t,d}$'s emerging from the progress of time and delay. In particular, the numbers $n_{t,d}$ are unknown when $d > T - t$. Hence, we shall denote the observed data as $\boldsymbol{n} = (n_{t,d} : (t,d) \in A_T^m)$, where

$$A_T^m = \{(t,d) : \max(T-m, 0) \leq t \leq T,\ 0 \leq d \leq \min(D, T-t)\}$$

is the right-angled trapezoidal observation region at time T when using a moving window of size m.

Denote by $N(t,T) = \sum_{d=0}^{\min(T-t,D)} n_{t,d}$ the number of cases which occurred on t and which are reported until time T. The aim of *nowcasting* is to predict the total number of cases

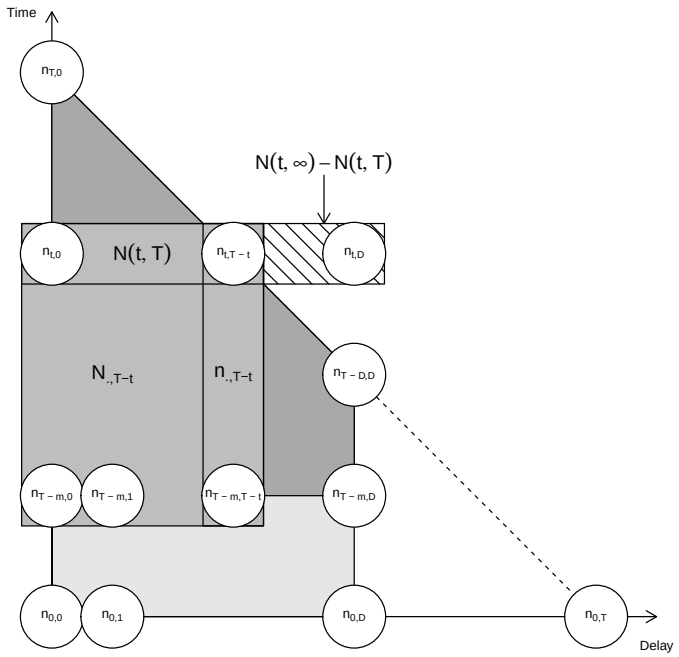

Figure 2: Reporting triangle at time T. The right-angled trapezoid spanned by $n_{0,0}, n_{T,0}, n_{T-D,D}$ and $n_{0,D}$ defines the available observations. Delays larger than D are not considered. The shaded box $N(t,\infty) - N(t,T)$ indicates the observations, which at time t are not yet available.

occurred on day t, $t = T - D, \ldots, T$, given the information available at time T. Formally, this number is $N(t, \infty) = \sum_{d=0}^{\infty} n_{t,d} = \sum_{d=0}^{D} n_{t,d}$ and consequently is the currently missing part $N(t, \infty) - N(t, T)$.

The reporting delay (in days) of a case occurring at time t follows a distribution with probability mass function (PMF) $f_t(d) = p_{t,d}$, $d = 0, \ldots, D$, where $\sum_{d=0}^{D} p_{t,d} = 1$. Considering time t and conditioning on the number of cases $N(t, T)$ observed by time T, the observed $n_{t,d}$'s in the reporting triangle due to the right-truncation have a multinomial distribution with size $N(t, \infty)$ and cell probabilities $p'_{t,d} = p_{t,d} / \sum_{i=0}^{\min(D, T-t)} p_{t,i}$ for $d = 0, \ldots, \min(D, T - t)$.

If one, as in Kalbfleisch and Lawless (1989) or Zeger et al. (1989), assumes an underlying inhomogeneous Poisson process on the occurrence of new cases s.t. $N(t, \infty) \sim \text{Po}(\lambda_t)$, $t = \max(0, T - m), \ldots, T$, the above conditional formulation of the $n_{t,d}$'s originating from multinomial sampling can be re-formulated as an unconditional likelihood corresponding to an incomplete contingency table

$$n_{t,d} \sim \text{Po}(\mu_{t,d}), \quad (t, d) \in A_T^m, \quad \text{where} \quad \log(\mu_{t,d}) = \log(\lambda_t) + \log(p_{t,d}). \tag{1}$$

Nowcasting can thus be divided into steps of determining the λ_t's, the sequence of $p_{t,d}$'s and finally predicting the unobserved $n_{t,d}$'s in order to compute the total $N(t, \infty)$.

We shall address the problem in Sect. 3 by first assuming a time-homogeneous delay distribution s.t. $p_{t,d} = p_d$ for all time points within the moving window. The starting point will be the work of Lawless (1994), which uses the conditional formulation to estimate the delay distribution without any assumptions about the distribution of the $n_{t,d}$'s. In a second step, we extend the delay distribution modelling in Sect. 4 to a discrete survival modelling context, which allows us to handle changes in the delay distribution over time. Here, the unconditional formulation becomes more natural to use, since the estimation of the epidemic curve and the delay can be formulated as a joint problem in the context of hierarchical log-linear modelling. This also allows us to impose smoothness constraints on the λ_t's by penalized splines.

3 Time homogeneous delay distribution

For the current "now" T, we will in this section assume time-homogeneity of the delay distribution within $t = \max(0, T - m), \ldots, T$, i.e we only take those cases into account, which occurred during the last $m+1$ days. In other words, we use as lower nodes of the right-angled trapezoid the observations $n_{T-m,0}$ and $n_{T-m,D}$ in Fig. 2. For the further derivation of the delay estimation we follow the notation in Lawless (1994) by introducing

$$n_{\cdot,d} = \sum_{t=\max(0,T-m)}^{T-d} n_{t,d}, \quad \text{and} \quad N_{\cdot,d} = \sum_{t=\max(0,T-m)}^{T-d} N(t, t+d),$$

i.e. $N_{\cdot,d}$ denotes the rectangle spanned by the observations $n_{\max(0,T-m),0}$, $n_{T-d,0}$, $n_{T-d,d}$ and $n_{\max(0,T-m),d}$. That is all regarded cases with delay less or equal to d. Furthermore, $n_{\cdot,d}$ denotes the right edge of this rectangle. This corresponds to all regarded cases with delay equal to d. Section B.1 in the Web Appendix contains an example calculation of these quantities for a specific reporting triangle.

Since at time T we observe a right-truncated version of the delay distribution, we define the reverse-time discrete hazard function (Lagakos et al., 1988; Kalbfleisch and Lawless, 1991;

Lawless, 1994),

$$g(d) = P(\text{delay} = d | \text{delay} \leq d) = \frac{f(d)}{F(d)}, \quad d = 0, 1, \ldots, D,$$

where $F(d)$ denotes the cumulative distribution function (CDF) of the delay distribution. We note that $g(0) = 1$. The CDF and PMF of the delay distribution are now given as

$$F(d) = \prod_{i=d+1}^{D} (1 - g(i)) \quad \text{and} \quad f(d) = F(d) - F(d-1) = g(d) \prod_{i=d+1}^{D} (1 - g(i)). \tag{2}$$

3.1 Frequentist nowcasting

For the reader's convenience in this section we recapitulate and discuss existing results on performing nowcasting for right-truncated delays. Lawless (1994) showed that the maximum likelihood estimate (MLE) for $g(d)$ when $T \geq D$ is $\hat{g}(d) = n_{\cdot,d}/N_{\cdot,d}$. The MLE for the delay distribution can now be obtained by plug-in of these estimates into (2). Note that if the window width m is very close to the maximum delay D, the MLE for the long delays may be based on few observations. Based on the MLE, Lawless (1994) presented the following nowcast procedure

$$\hat{N}(t, \infty) = \frac{N(t, T)}{\hat{F}(T - t)},$$

which can be linked to inverse probability weighting. This link also provides solutions for the case when $\widehat{F}(T - t)$ is zero (Elliot and Little, 1999). If $N(t, T) = 0$, i.e. no events are reported to have occurred at t by time T, then $\hat{N}(t, \infty)$ equals zero. Otherwise, the predictive distribution can be computed based on an asymptotic normal approximation for the prediction error. This is done by computing the variance of $Z_t = N(t, \infty) - \hat{N}(t, \infty)$. From this, a $(1-\alpha)\cdot 100\%$ prediction interval for $N(t, \infty)$ is $\hat{N}(t, \infty) \pm z_{1-\alpha/2} \widehat{\text{Var}}(Z_t)^{\frac{1}{2}}$, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. We shall obtain a discrete predictive distribution for $N(t, \infty)$ via discretizing this Gaussian predictive distribution by taking the difference of the CDF evaluated at the integers and attribute all mass below $N(t, T)$ to $N(t, T)$.

3.2 Bayesian nowcasting

As a competitor to the above method we propose a novel Bayesian hierarchical model to directly address the combination of delay estimation and count data nature of the forecast. In particular, we will show that the generalized Dirichlet (GD) distribution is a conjugate prior-posterior distribution for the reversed delay under right-truncated multinomial sampling.

The density of the GD($\boldsymbol{\alpha}, \boldsymbol{\beta}$) distribution (Wong, 1998) with parameters $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{D-1})'$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{D-1})'$ is proportional to

$$f(\text{rev}(\boldsymbol{p})) \propto \prod_{i=0}^{D-1} p_{D-i}^{\alpha_i - 1} (1 - p_D - \ldots - p_{D-i})^{\gamma_i},$$

where the probabilities $p_D, \ldots, p_1$ and $p_0 = 1 - p_D - \cdots - p_1$ are all non-negative and sum to one. Here, $\gamma_i = \beta_i - \alpha_{i+1} - \beta_{i+1}$ for $i = 0, \ldots, D-2$ and $\gamma_{D-1} = \beta_{D-1} - 1$. In the above,

rev$(\cdot)$ denotes the reverse delay order, i.e. where delay d is mapped to the new scale $D-d$ in analogy with the reverse time hazard function.

Property 3 in the Web appendix shows that the posterior of rev$(\boldsymbol{p})$ under right-truncated multinomial sampling of Sect. 2 is again a GD distribution with parameters $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$,

$$\alpha_i^* = \alpha_i + n_{.,D-i}, \quad \text{and} \quad \beta_i^* = \beta_i + N_{.,D-i} - n_{.,D-i}, \quad \text{for } 0 \leq i \leq D-1.$$

With the above generalized Dirichlet posterior the marginal posterior expectation for $p_{D-i}, 0 \leq i \leq D-1$, when using the improper prior $\boldsymbol{\alpha} = \boldsymbol{\beta} = \mathbf{0}$ can be shown to be equal to the ML estimate defined in the previous section (see Web Appendix).

Note that the posterior in this case is only proper if $n_{.,d} > 0$ and $N_{.,d} > 0$ for all $d = 0, \ldots, D$. In practice, we shall use the informative prior $\boldsymbol{\alpha} = (\kappa, \kappa, \ldots, \kappa)'$, $\beta_{D-1} = \kappa$ and $\beta_i = \alpha_{i+1} + \beta_{i+1}$, $i = 0, \ldots, D-2$, where $\kappa > 0$ is a known constant describing the prior concentration. In this case the GD prior reduces to a symmetric Dirichlet distribution with concentration parameter κ (see Wong (1998) or Web Appendix). A property of this prior is that $\mathrm{E}(p_{D-i}) = 1/(D+1)$ for $i = 0, \ldots, D$. Note that if one uses a Dirichlet prior and pretends that the column sums in the reporting triangle originate from multinomial sampling, then the posterior is again Dirichlet (and hence also GD) and represents a Bayesian estimate ignoring right-truncation (see Web Appendix).

Assume now that, by conjugate prior-posterior updating, we have obtained an estimate of the delay distribution, i.e. rev$(\boldsymbol{p}) \sim \mathrm{GD}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. In order to predict the unknown $N(t,\infty), t = T-D, \ldots, T$, from the incomplete $N(t,T)$ we assume the following model hierarchy for $t = T-D, \ldots, T$

$$\begin{aligned} \lambda_t &\sim \mathrm{Ga}(\alpha_\lambda, \beta_\lambda), \\ N(t,\infty)|\lambda_t &\sim \mathrm{Po}(\lambda_t), \\ N(t,T)|N(t,\infty), q_{T-t} &\sim \mathrm{Bin}(N(t,\infty), q_{T-t}), \end{aligned}$$

where $q_{T-t} = \sum_{d=0}^{T-t} p_d$ is the proportion reported within a delay of $T-t$ days and $\alpha_\lambda, \beta_\lambda > 0$ are known constants. In this hierarchy, the marginal (prior) distribution of $N(t,\infty)$ is negative binomial with mean $\mu_\lambda = \alpha_\lambda \beta_\lambda$ and variance $\mu_\lambda + \mu_\lambda^2/\alpha_\lambda$. Furthermore, given q_{T-t} the distribution of $N(t,T)$ is compound binomial-negative binomial. The marginal posterior for $N_t = N(t,\infty)$, $t = T-D, \ldots, T$, is

$$f(N_t|N(t,T)) = \int_0^1 f(N_t|q_{T-t}, N(t,T)) f(q_{T-t}|N(t,T)) dq_{T-t}, \tag{3}$$

where by application of Bayes' theorem and using $p = 1/(1+\beta_\lambda)$

$$\begin{aligned} f(N_t|q_{T-t}, N(t,T)) &= \binom{N_t}{N(t,T)} q_{T-t}^{N(t,T)} (1-q_{T-t})^{N_t - N(t,T)} \binom{N_t + \alpha_\lambda - 1}{N_t} p^{\alpha_\lambda} (1-p)^{N_t}, \\ &\propto \frac{\Gamma(N_t + \alpha_\lambda)}{\Gamma(N_t - N(t,T) + 1)} \left(\frac{(1-q_{T-t})\beta_\lambda}{1+\beta_\lambda} \right)^{N_t}, \end{aligned} \tag{4}$$

for $N_t \geq N(t,T)$ and zero otherwise. Note that $f(q_{T-t}|N(t,T))$ in (3) is not available in analytic form, but from the previous developments we know that rev$(\boldsymbol{p})|N(t,T) \sim \mathrm{GD}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$. We hence solve the integration in (3) by Monte Carlo sampling jointly q_{T-t} for all $t = T-D, \ldots, T$ using the following algorithm.

(a) For $k = 1, \ldots, K$:

(1) Draw $\text{rev}(\boldsymbol{p}^{(k)}) \sim \text{GD}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ by the algorithm of Wong (1998) and calculate $q_{T-t}^{(k)} = \sum_{d=0}^{T-t} p_d^{(k)}$ for $T - D \leq t \leq T$.

(2) Calculate the non-normalized density $\tilde{f}_{n,t}^{(k)} = f(N(t,T)|N_t = n, q_{T-t}^{(k)}) f(N_t = n|\lambda_t) f(\lambda_t)$ for $T - D \leq t \leq T$ and $n = 0, 1, \ldots, N_{\max}$, where $N_{\max}$ is sufficiently large.

(3) Set $f_{n,t}^{(k)} = \tilde{f}_{n,t}^{(k)} / c_t^{(k)}$, where $c_t^{(k)} = \sum_{n=0}^{N_{\max}} \tilde{f}_{n,t}^{(k)}$ is the normalization constant.

(b) Approximate $f(N_t|N(t,T)) \approx K^{-1} \sum_{k=1}^{K} f_{n,t}^{(k)}$ for $n = 0, 1, \ldots, N_{\max}$ and $t - D \leq t \leq T$.

Since the Monte Carlo sampling in (b) is based on entire probability vectors and not just samples, only a small number of K, say 100 or 1000, is needed to obtain accurate results for the multivariate integration. Altogether, our Monte Carlo based procedure is fast, accurate and easy to use in practice.

4 Joint Bayesian modelling of epidemic curve and delay distribution

In outbreak situations there is often prior knowledge about the shape of the epidemic curve (here measured as daily number of hospitalizations): typically, one expects the curve to initially increase, level off and then decrease. An advantage of such shape assumptions is that they can help to improve the estimation of the delay distribution and hence the nowcasts (Kalbfleisch and Lawless, 1989). However, for foodborne outbreaks such shape assumptions are only partially applicable: the epidemic curve is very much dominated by how the contaminated food item is distributed, its shell life and consumer awareness. Person-to-person transmission often only plays a minor role. Hence, one of the concerns during the STEC outbreak was, if several waves could occur. As a consequence, our modelling of the epidemic curve is restricted to a semi-parametric approach imposing some degree of smoothness.

4.1 Epidemic curve

In Sect. 3.2, we used $\lambda_t \overset{iid}{\sim} \mathcal{G}a(\alpha_\lambda, \beta_\lambda)$ as prior for the expected number of cases. Instead we now model $\log(\lambda_t)$ as a quadratic spline based on the TP basis with knots placed at $\{\kappa_1, \ldots, \kappa_L\}$ to describe the evolution in λ_t. We place $\min(\lfloor T/6 \rfloor, 20)$ knots evenly between time zero and T. In order to avoid over-extrapolations at the end, where uncertainty is large, we remove any knots which are between time $T - D/2$ and T. Then,

$$\log(\lambda_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \sum_{i=1}^{L} b_i \max(t - \kappa_i, 0)^2. \tag{5}$$

The first part of (5) is a global 2nd order polynomial reflecting that the epidemic curve initially rises, levels off and then decreases again. The second part contains terms representing the deviation from this global polynomial by having the 2nd order terms change at each knot. Automatic smoothing can be obtained by treating the b_i's as independent random effects, i.e. $(b_1, \ldots, b_L)' \sim N(\boldsymbol{0}, \sigma_b^2 \boldsymbol{I})$ with the amount of smoothing being proportional to $1/\sigma_b^2$. By setting up design matrices $\boldsymbol{X} = [\boldsymbol{1}, \boldsymbol{t}, \boldsymbol{t}^2]$ and $\boldsymbol{Z} = [\max(\boldsymbol{t} - \kappa_1, 0)^2, \ldots, \max(\boldsymbol{t} - \kappa_L, 0)^2]$ it becomes obvious that this corresponds to a Poisson mixed model, which can be handled in a Bayesian framework by appropriate specification of priors (Fahrmeir et al., 2013, Chapter 8.1.9).

4.2 Delay distribution modelling

In Sect. 3 we assumed the same delay distribution for all times t. To generalize this, we instead use a discrete time survival regression model to describe the truncated delay distribution. We shall use a model which combines the non-parametric-like feature of the Dirichlet distribution (i.e. a parameter for each delay time) with parametric features for modelling changes in the distribution. Specifically, we model the discrete time hazard $h_{t,d} = P(\text{delay} = d|\text{delay} \geq d, \boldsymbol{W}_{t,d})$, $d = 0, \ldots, D$, for a case occurring at time t as

$$\text{logit}(h_{t,d}) = \gamma_d + \boldsymbol{W}'_{t,d}\boldsymbol{\eta}, \quad d = 0, \ldots, D-1 \quad \text{and} \quad h_{t,D} = 1.$$

Here, $\boldsymbol{W}_{t,d}$ denotes a $p \times 1$ vector of covariates depending on time and delay and $\boldsymbol{\eta}$ are the corresponding covariate effects. The corresponding delay distribution is then computed as

$$p_{t,d} = \left(1 - \sum_{i=0}^{d-1} p_{t,i}\right) h_{t,d}, \quad d = 0, \ldots, D. \tag{6}$$

As an example: On 23 May 2011 there was an intervention towards faster reporting, where local and state health authorities were asked to communicate suspected HUS cases directly and without the otherwise necessary quality checking. This situation could now be represented by a single change-point s.t. $\boldsymbol{W}_{t,d} = \mathbb{1}(t + d \geq \texttt{2011-05-23})$. Extension of the above modelling to include day-of-the-week effects or case specific covariates, e.g. age or county of residence, is immediate.

4.3 Implementation of the hierarchical Bayesian nowcasting model

Multivariate normal priors on $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ for the model (6) were combined with the spline model in (5) resulting in a Bayesian hierarchical Poisson model for nowcasting based on (1). Inference per T was done using MCMC as implemented in JAGS (Plummer, 2003) using the `runjags` package (Denwood, 2013). Three separate chains were run using parallel computing on three CPU kernels simultaneously. To obtain better convergence the time covariate was centered as $t' = t - T/2$ in the TP spline. Furthermore, the number of γ's in the discrete survival model was halved by using an alternative parameter vector $\boldsymbol{\gamma}'$ of length $\lfloor (D-1)/2+1 \rfloor$ and then assuming $\gamma_d = \gamma'_{\lfloor d/2+1 \rfloor}$ in order to stabilize and speed up the computations.

5 Evaluating the Nowcasts

We assess the quality of the nowcasts by using proper scoring rules for count data (Czado et al., 2009) while comparing with the complete data at the end of the outbreak. Such scoring rules have become the standard the way for evaluating probabilistic forecasts in count data time series applications, because they address both the aspect of calibration and sharpness – see, e.g., Paul and Held (2011) for an application in infectious disease modelling. Here, calibration refers to the consistency between the distributional forecast and the observation that materializes. Sharpness refers to the concentration of the forecast distribution.

We denote by $\mathcal{T}$ the set of days at which to do the nowcasting. For each $T \in \mathcal{T}$ we consider the k time points $\mathfrak{T}_T = \{T - k - l + 1, \ldots, T - l\}$, where l is the first lag for which the forecasts are considered reliable and $k - l \leq D$ is the last lag at which nowcasts are considered relevant. In our case, we will use $k = 10$ and $l = 3$ since nowcasts for delays between 0-2 days were too dominated by chance to be communicable.

Considering a specific time T and a single $t \in \mathfrak{T}_T$, let P_t^T be the predictive distribution for time $t = T - d$ based on the information available at T and with $N(t, \infty)$ being the true. In the evaluations we will use the *logarithmic score* (logS)

$$\text{logS}(P_t^T, N(t,\infty)) = -\log(f_{P_t^T}(N(t,\infty))),$$

where $f_{P_t^T}(\cdot)$ is the PMF of the predictive distribution P_t^T. This rule is very intuitive – the higher the probability of the forecast distribution for the actual observed value the better. However, the rule is local, because no credit is given for high forecast probabilities near the true value. The *ranked probability score* (RPS)

$$\text{RPS}(P_t^T, N(t,\infty)) = \sum_{k=0}^{N_{\max}} \Big(F_t^T(N(t,\infty)) - \mathbb{1}(N(t,\infty) \leq k)\Big)^2$$

is a less local proper scoring rule, since comparison between the CDF of the predictive distribution, F_t^T, and the empirical one-value CDF happens over the entire support. If the predictive distribution is calculated by tabulating samples, e.g., MCMC output, another disadvantage of logS is that it becomes infinity, if no samples are equal to the true value. This can quickly happen, if the true predictive probability is low.

The output of individual scoring rules can be combined by averaging. For example, the mean score of a specific nowcast method at lag d is

$$\overline{SR}(d) = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} SR(P_t^T, N(t,\infty)),$$

where SR is the scoring rule. An alternative is to perform the summation over the lags instead of the elements in $\mathcal{T}$; one then obtains a mean-score per T, i.e. $\overline{SR}(T)$. In some cases one wants to average over all scores, i.e. calculate $\overline{SR} = \sum_{d=l}^{k+l-1} \overline{SR}(d)/k$. Note that in such overall mean calculations, the same true value $N(t, \infty)$ enters once for each $\mathfrak{T}_T$, where t is included, i.e. up to k times. Furthermore, one should be aware that averaging over all lags means that one considers nowcasts far away from "now" to be just as important to casts about days close to "now". In practice, one may might be more interested in getting good predictions about the near past.

6 Nowcasting during the STEC O104:H4 outbreak

We evaluate the quality of different nowcast procedures during the outbreak based on the date of hospitalization for HUS, because this proved to be a reliable and specific indicator for outbreak cases. Evaluation starts 2011-06-02, which was the day of the first nowcast produced during the outbreak (Robert Koch Institute, 2011), and continues to 2011-07-04, where the outbreak was officially declared to be over. For each day, all currently available cases are included for the estimation of the delay distribution. Delays up to $D = 15$ days are considered – if any of the available cases have a longer delay, we set their delay to the upper limit of 15 days. This occurred for 11.9% of the 658 HUS cases with an available hospitalization date. Note: The specific choice of $D = 15$ is motivated by a combination of a median delay of 5-6 days, estimation robustness and a lack of relevance in adjusting total hospitalization counts too far back. However, information about cases with very long delays is not lost in the delay distribution estimation, since the last category corresponds to 'Delay $\geq D$'. If the interest had been in adjusting counts further back, one could have used a larger

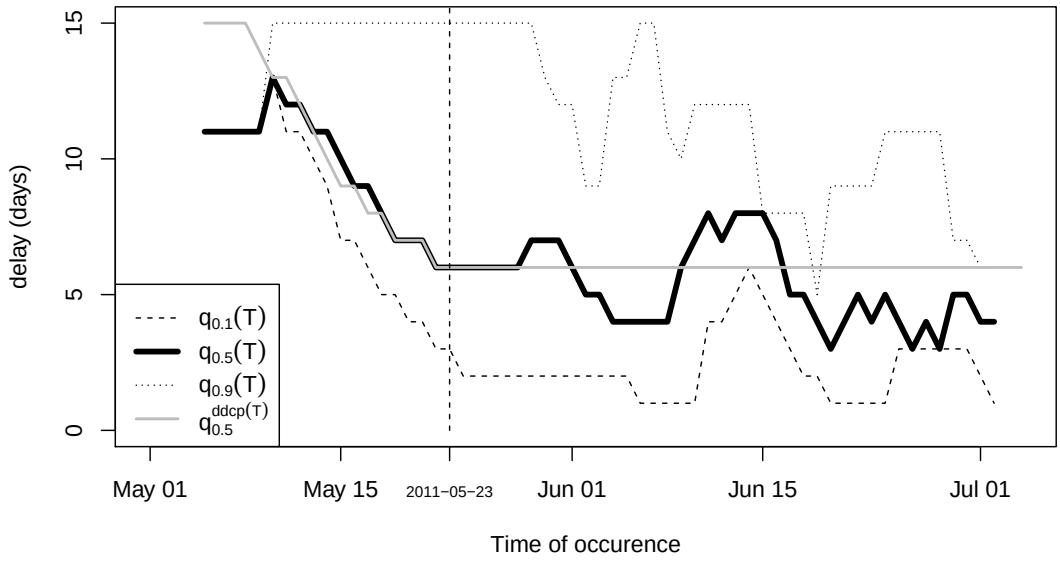

Figure 3: Median delay (as well as 10% and 90% quantiles) of the smoothed empirical delay distribution as a function of time of occurence. Also shown is the resulting median when the posterior mean of the single change-point model (fitted on 2011-07-04) is inserted into (6).

D, but would then have had to categorize delays in courser time units (e.g. as in Sect. 4.3) to ensure stability of the estimation.

Retrospectively, based on all cases, it was possible to assess the delay distribution $(p_{t,d})_{d=0}^{D}$ for each day t by tabulating the delays for all cases occuring at time t in the final reporting triangle. Figure 3 contains the daily median of the smoothed empirical delay distribution obtained from tabulating all cases occuring within a moving window of $t-2,\ldots,t+2$.

We observe a distinct delay reduction due to the intervention on 23 May 2011. This reduction is well captured by the changepoint model. The plot also reveals some unexpected variation in the delay distribution between 10-20 June 2011. However, since these estimates are based on just a few cases and no apriori explanations exist for these changes, we decided to not add extra changepoints in the delay distribution model. Daily estimates of the delay distribution from the bayes.trunc.ddcp procedure are found as Web Animation 2. Furthermore, daily ML estimates of the delay distribution with and without right-truncation adjustment are illustrated in Fig. 2 of the Web Appendix.

As an example of how the delay estimates are translated into nowcasts for the number of hospitalizations, Fig. 4 shows the predictive distribution of $N_T(t,\infty)$ when T=2011-06-02 and t=2011-05-28 for the Lawless approach and the three Bayesian procedures: GD-Negbin Bayes ignoring truncation (abbreviated: bayes.notrunc), truncation adjusted GD-NegBin Bayes (bayes.trunc) and the hierarchical Bayes procedure (bayes.trunc.ddcp) with delay distribution changepoint and smoothed epidemic curve. Furthermore, we chose $N_{\max}=300$ in the PMF computation of the non-MCMC Bayes procedures.

An animation showing the median and 95% credibility interval of the nowcasts obtained with the bayes.trunc.ddcp procedure is shown in Web Animation 3. In the animation, the delay distribution is re-estimated for each day based on all currently available observations.

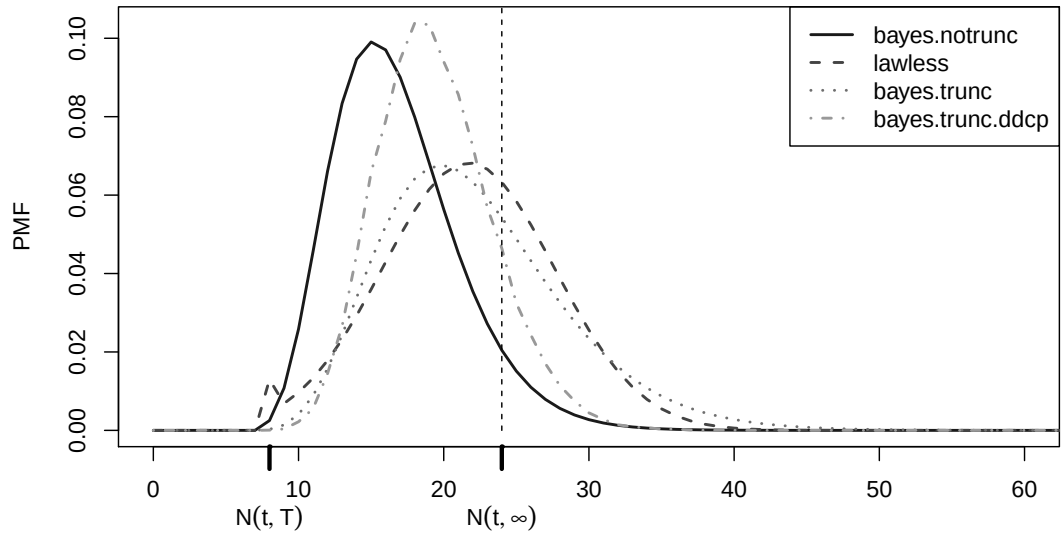

Figure 4: Predictive distribution of the number of hospitalizations $N(t, \infty)$ when T=2011-06-02 and t=2011-05-28, i.e. the considered delay is $T - t = 5$ days. The true number of hospitalizations is $N(t, \infty) = 24$.

	bayes.notrunc	lawless	bayes.trunc	bayes.trunc.ddcp	unif
logS	2.28	Inf	2.02	Inf	5.69
RPS	1.79	15.23	1.77	1.39	95.10
dist.median	2.37	2.50	2.51	1.87	143.94
outside.ci	0.07	0.16	0.05	0.09	0.84

Table 1: Mean score for the period from `2011-06-02` and `2011-07-04`.

The animation also contains a plot of the posterior median of λ_t and an equal tailed 95% credibility interval, which gives a good indication of the current epidemic trend (up, stable or down) and its associated uncertainty.

6.1 Nowcast comparisons using scoring rules

The following procedures are compared: the truncation adjusted Lawless (1994) procedure and the three previously described Bayes procedures (bayes.notrunc, bayes.trunc and bayes.notrunc.ddcp). In order to form a baseline for our comparisons we also assess the behaviour of the discrete uniform (unif) on $N(t, T), \ldots, N_{\max}$ as predictive distribution. We evaluate the procedures between 2011-06-02 and 2011-07-04 using $l = 3$ and $k = 10$. Altogether, $k \cdot |\mathcal{T}| = 10 \cdot 33 = 330$ nowcasts are computed.

To quantify the comparisons we use four scores: the logarithmic-score and RPS as proper scoring rules together with the more heuristic measures: median of the absolute deviations and proportion of times the observed value lay outside an equal-tailed 95% prediction interval. The mean overall score of all 330 nowcasts is given in Tab 1.

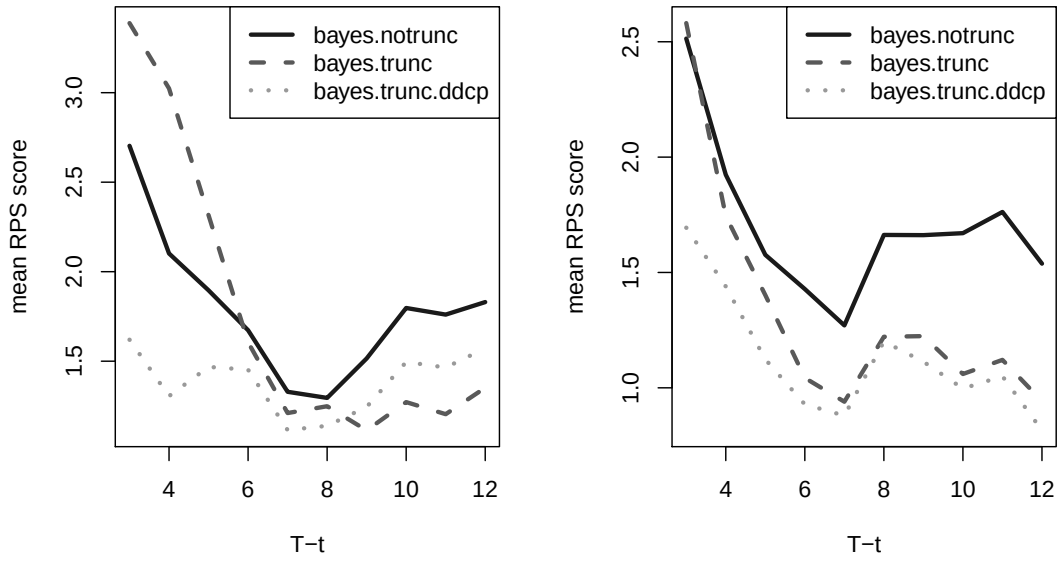

Figure 5: Mean RPS Score by lag. Left panel shows the results for the actual data. For comparison, the right panel shows results for simulated data with stationary delay distribution.

Based on RPS, bayes.trunc.ddcp has the overall best performance. It also becomes clear that the Lawless procedure is not performing too well for the HUS outbreak data. Partly, this is caused by the fact that an observation of zero counts leads to a point mass forecast distribution at zero. As a consequence, the lawless procedure has infinite mean log-score. Altogether, the approximate Gaussian distribution of the predictive appears unwarranted in the small count data setting. The bayes.trunc.ddcp procedure also has an infinite mean log-score, because the predictive PMF is computed by tabulating 30,000 MCMC samples at each instance. Hence, in some situations where the probability is low the resulting PMF is zero, which leads to an infinite log-score when evaluated at this value.

An investigation of the effect of the lag $T - t$ on the RPS score is shown in the left panel of Fig. 5. To further interpret this figure, we show corresponding results in simulated data, where the delay distribution is time-constant. We obtain this simulated data by using the actual event time for each case, but simulate report time according to a PMF equal to the delay distribution of all 658 cases (delays >15 days were set to 15 days). The right panel in Fig. 5 shows the corresponding results. We note that for the real outbreak data, the procedure ignoring truncation perform better for time points closer to "now" (typically these being the more interesting ones to nowcast) than the purely right truncation adjusting method. For the simulated data the right-truncation adjusting procedures performs better for all lags, which is further indication that the delay distribution changed during the outbreak and that the bias of the non-adjusting procedures (too much weight on the short delays) coincides with the real delays getting shorter. Except for long lags in the actual data does bayes.trunc.ddcp outperform the two other procedures; for the simulated data this result is an emphasis of the effect of the epidemic curve on the nowcasts.

As a final comparison, Fig. 6 shows the mean RPS score as a function of the nowcasted time points in $\mathcal{T}$. We observe high values of the mean-score for the early time points, where

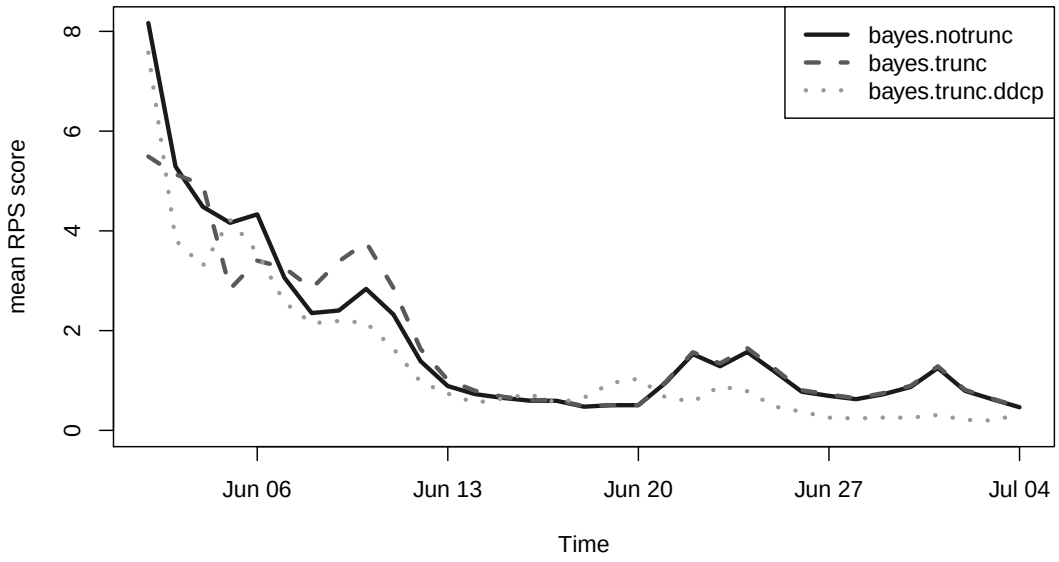

Figure 6: Mean-score as a function of time T, where the nowcasting is performed.

the number of cases is high. Altogether, when averaging over the 10 included lags, the bayes.trunc.ddcp procedure appears to perform best. However, early in the outbreak the procedure is not superior – even though it is the only procedure adjusting for the change in delay distribution. A closer examination reveals that the combination of a flexible delay distribution model and the 2nd order polynomial leads to over-extrapolation in this situation. Stronger priors might be a way to obtain better results this situation. Note also that the further scoring gets away from the 23 May 2011, the more are differences between the bayes.trunc and bayes.trunc.ddcp procedures due to the penalized TP spline acting on the epidemic curve. Altogether, it appears that once the peak has been reached the hierachical model works very well, while in the early phases it might contain too much flexibility.

In conclusion, ignoring truncation for the outbreak data is an ad-hoc way to compensate for the fact that, early in the outbreak, delays reduced significantly. However, right-truncation adjusted procedures are to be recommended as proper adjustment methods. A simple but robust way to handle time varying delay distributions with the methods from Sect. 3 is to use a moving window in the delay distribution estimation. Figures 4 and 5 in the Web Appendix contain the results of the daily updated delay distribution estimate when using a moving window of 25 days. For comparison, Tab. 2 in the Web Appendix contains the mean scores when using a moving-window procedure of 25 days and Fig. 6 in the Web Appendix shows the mean scores by lag. A small improvement is noticed for all Bayes procedures. However, a moving window based procedure remains a compromise between including only recent events and having enough events to obtain a reliable estimation. Altogether, the reporting dynamics of the STEC outbreak emphasizes the need to handle time-varying reporting delay distributions – either using simple moving window approaches or more complex regression models for the delay distribution. Furthermore, the results show that imposing a smoothness restriction on the epidemic curve as done in Sect. 4 helps improve the nowcasts. At the early stages of the outbreak, when only few observations are available, such flexible models are,

however, difficult to identify and the interplay between epidemic curve component and delay distribution component in the estimation is non-trivial.

7 Discussion

Real-time tracking of the epidemic curve during the STEC/HUS outbreak was an important component of the outbreak management. The results of the nowcasts were circulated amongst the stakeholders at the RKI on a daily basis and were hence part of the body of evidence used to assess epidemiological trends each day. In case communication of a more qualitative output for the epidemic trend is required, an additional decision procedure could categorize the nowcasts into, e.g., trend states "up", "stable" and "down". Such answers are immediate from the penalized TP spline for $\log(\lambda_t)$ as visualized in the Web Animation 3. Because the outbreak was a time-limited event, it was in retrospect possible for us to determine the true number of daily hospitalizations due to HUS. Having the true number of hospitalizations allowed a unique evaluation of the quality of the nowcasts by proper scoring rules.

Our comparisons showed that the proposed Bayesian nowcast procedures, which took the count data nature into account, provided better results than the discretized Lawless procedure.

Within the Bayesian procedures, addressing right-truncation under the assumption of a time homogeneous delay distribution only provided small improvements for the STEC/HUS outbreak. The main reason is that the actual delay pattern was time-varying, because of intervention measures to ensure faster reporting and other complex factors, e.g. awareness of hospital doctors towards HUS and its reporting, resources at local health institutions, media attention and case definition. Altmann et al. (2011) contains a more detailed analysis of the reporting delay during the outbreak. Since the intervention towards faster reporting occurred at an early stage of the outbreak, a moving window based approach is a way to address the problem. The model based approach reflecting knowledge about the intervention, however, showed better results and appears less ad-hoc. The Poisson formulation of the nowcasting problem also makes it obvious that a frequentist alternative could have been to fit a generalized additive model containing a penalized spline for the trend and a factor for the delay (Wood, 2006). However, the hierarchical Bayes approach is more flexible, because it allows the combination of complex models for both delay and epidemic curve. Furthermore, predictive distributions are immediately available due to the data-augmenting nature of the Bayes approach.

The evaluations based on scoring rules depend on the specific time-period investigated, i.e. it makes a difference whether the early outbreak with many cases and fluctuating reporting patterns or the late outbreak period with few cases and less dramatic changes is investigated. Our choice of evaluation period for the scoring was based on an interest of what method – in retrospect – would have been the best way to address the nowcasting problem as posed by the circumstances of the outbreak then. As easy as it is to model details in hindsight, we underline that during outbreaks *consistency is more important than accuracy* (Cowden, 2010) and hence robustness is often more important than the best fit. To promote application of nowcasting procedures in future outbreak situations or even non-infectious disease applications, the proposed methods are available in the R package `surveillance` (Höhle, 2007) as function `nowcast`.

7.1 Supplementary Materials

Web Appendices, Tables, Figures and Animations referenced in the Sections 1, 3 and 6 are available with this paper at the Biometrics website on Wiley Online Library. Furthermore, a R file for analysing the data corresponding to Fig. 1 of the Web appendix is available from this website[1].

7.2 Acknowledgments

We thank Leonhard Held for fruitful discussions and helpful suggestions at the initial stages of our work. Thanks also goes to Doris Altmann for preparing the SQL query, which extracted the necessary data for our analysis from the SurvNet@RKI database. Comments and suggestions by the associate editor and two anonymous reviewers helped improve the manuscript substantially.

References

Altmann, M., Wadl, M., Altmann, D., Benzler, J., Eckmanns, T., Krause, G., Spode, A., and an der Heiden, M. (2011). Timeliness of surveillance during outbreak of shiga toxin–producing Escherichia coli infection, Germany, 2011. *Emerging Infectious Diseases*, 17(10):1906–1909.

Brookmeyer, R. and Damiano, A. (1989). Statistical methods for short-term projections of AIDS incidence. *Stat Med*, 8(1):23–34.

Buchholz, U., Bernard, H., Werber, D., Böhmer, M. M., Remschmidt, C., Wilking, H., Deleré, Y., a. d. Heiden, M., Adlhoch, C., Dreesman, J., Ehlers, J., Ethelberg, S., Faber, M., Frank, C., Fricke, G., Greiner, M., Höhle, M., Ivarsson, S., Jark, U., Kirchner, M., Koch, J., Krause, G., Luber, P., Rosner, B., Stark, K., and Kühne, M. (2011). German outbreak of Escherichia coli O104:H4 associated with sprouts. *New England Journal of Medicine*, 365:1763–1770.

Cowden, J. M. (2010). Some haphazard aphorisms for epidemiology and life. *Emerging Infectious Diseases*, 16(1):174–177.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65:1254–1261.

Denwood, M. J. (2013). runjags: An R package providing interface utilities, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*. In review.

Donker, T., van Boven, M., van Ballegooijen, W. M., Van't Klooster, T. M., Wielders, C. C., and Wallinga, J. (2011). Nowcasting pandemic influenza A/H1N1 2009 hospitalizations in the Netherlands. *European Journal of Epidemiology*, 26(3):195–201.

Elliot, M. R. and Little, R. J. A. (1999). Weight trimming in a random effects model framework. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 365–370.

[1]Since the surveillance package is continuously developed, the help of the function `nowcast` might contain more up-to-date information on how to use it.

England, P. D. and Verrall, R. J. (2002). Stochastic claims reserving in general insurance. Sessional meeting paper, Institute of Actuaries and Faculty of Actuaries.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer.

Frank, C., Werber, D., Cramer, J. P., Askar, M., Faber, M., an der Heiden, M., Bernard, H., Fruth, A., Prager, R., Spode, A., Wadl, M., Zoufaly, A., Jordan, S., Kemper, M. J., Follin, P., Müller, L., King, L. A., Rosner, B., Buchholz, U., Stark, K., and Krause, G. (2011). Epidemic profile of shiga-toxin–producing Escherichia coli O104:H4 outbreak in Germany. *New England Journal of Medicine*, 365(19):1771–1780.

Green, H. K., Andrews, N. J., Bickler, G., and Pebody, R. G. (2012). Rapid estimation of excess mortality: nowcasting during the heatwave alert in England and Wales in June 2011. *Journal of Epidemiology and Community Health*, 66(10):866–868.

Höhle, M. (2007). surveillance: An R package for the surveillance of infectious diseases. *Computational Statistics*, 22(4):571–582.

Kalbfleisch, J. D. and Lawless, J. F. (1989). Inference based on retrospective ascertaintment: An analysis of the data on tranfusion related AIDS. *Journal of the American Statistical Association*, 84(406):360–372.

Kalbfleisch, J. D. and Lawless, J. F. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica*, 1:19–32.

Lagakos, S. W., Barraj, L. M., and De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika*, 75(3):515–523.

Lawless, J. F. (1994). Adjustments for reporting delays and the prediction of occurred but not reported events. *The Canadian Journal of Statistics*, 22(1):15–31.

Lin, H., Yip, P. S. F., and Huggins, R. M. (2008). A double-nonparametric procedure for estimating the number of delay-reported cases. *Statistics in Medicine*, 27:3325–3339.

Midthune, D. N., Fay, M. P., Clegg, L. X., and Feuer, E. J. (2005). Modeling reporting delays and reporting corrections in cancer registry data. *Journal of the American Statistical Association*, 100(469):61–70.

Nunes, B., Natário, I., and Lucília Carvalho, M. (2013). Nowcasting influenza epidemics using non-homogeneous hidden Markov models. *Statistics in Medicine*, 32(15):2643–2660.

Paul, M. and Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine*, 30(10):1118–1136.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Hornik, K., Leisch, F., and Zeileis, A., editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Robert Koch Institute (2011). Final presentation and evaluation of epidemiological findings in the EHEC O104:H4 outbreak, Germany, 2011. Technical report, Robert Koch Institute.

Wong, T.-T. (1998). Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation*, 97:165–181.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.

Zeger, S. L., See, L. C., and Diggle, P. J. (1989). Statistical methods for monitoring the AIDS epidemic. *Stat Med*, 8(1):3–21.