

L04 Latencies and Delays¹

Michael Höhle¹

¹Department of Mathematics, Stockholm University, Sweden



m_hoehle

STA427 FS2021

Statistical Methods in Infectious Disease Epidemiology
Epidemiology, Biostatistics and Prevention Institute
University of Zurich, Switzerland



University of
Zurich^{UZH}

Outline

- 1 Back-calculation method
 - Back-projection for the 2011 STEC/HUS outbreak
 - Discussion and Extensions
- 2 Nowcasting

Overview

- 1 Back-calculation method
 - Back-projection for the 2011 STEC/HUS outbreak
 - Discussion and Extensions

- 2 Nowcasting

STEC/HUS Outbreak in Germany 2011 (1)

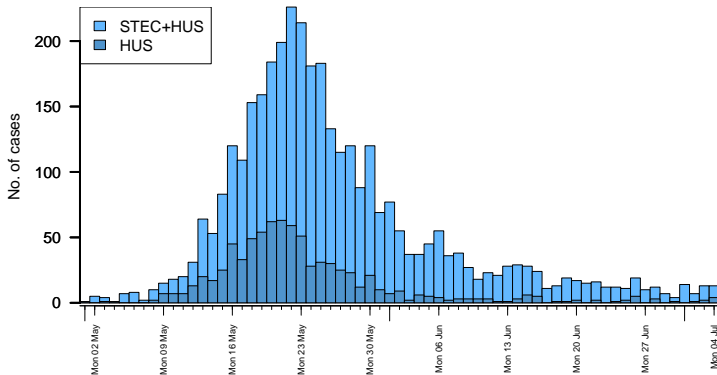
- Outbreak of Shiga toxin-producing *E. coli* (STEC) O104:H4 in Germany May–July 2011 associated with sprouts

	STEC	HUS
N (% of total)	2987 (78)	855 (22)
Median age (years)	46	42
Female (%)	58	68
Deaths	18	35
Case-fatality-ratio (%)	0.6	4.1

- Hemolytic-uremic syndrome (HUS) is a disease characterized by hemolytic anemia, thrombocytopenia and acute kidney failure.
- HUS can be a complication of an STEC infection.
- Onset of HUS occurs a median of 5 days (IQR: 4–7) days after onset of the STEC related diarrhea.

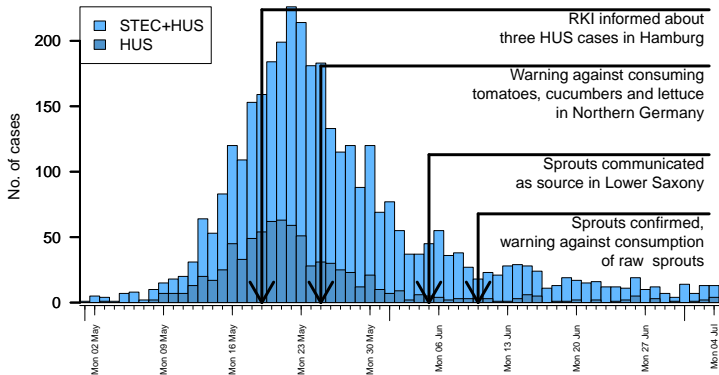
STEC/HUS Outbreak in Germany 2011 (2)

- Retrospective curve illustrating the onset of diarrhea of confirmed patients per day (where available: STEC 2715, HUS 783)



STEC/HUS Outbreak in Germany 2011 (2)

- Retrospective curve illustrating the onset of diarrhea of confirmed patients per day (where available: STEC 2715, HUS 783)



Example: STEC/HUS Outbreak in Germany 2011 (3)

- However, during the outbreak the situation is not as clear.
- Incubation period and reporting delays complicate real-time tracking of key indicators for detecting epidemic trends.
- Illustration: Day of hospitalization of HUS cases and the day the HUS case arrives at the RKI.

[Animated curve of reporting delay of HUS cases]

Focus on implication of time lags

Time lags during the STEC outbreak, e.g.,

- the delay between exposure to the disease and onset of diarrhea in cases
- the inherent reporting delay present in any public health surveillance system

Goal of back-projection:

Infer exposure times of HUS patients from the retrospective epidemic curve of diarrhea onsets in order to reconstruct the infection curve.

Goal of nowcasting:

Extrapolate currently available counts by taking the reporting delay from the past into account. Add uncertainty indication to this extrapolation.

Outline

- 1 Back-calculation method
 - Back-projection for the 2011 STEC/HUS outbreak
 - Discussion and Extensions
- 2 Nowcasting

Motivation for back-projection

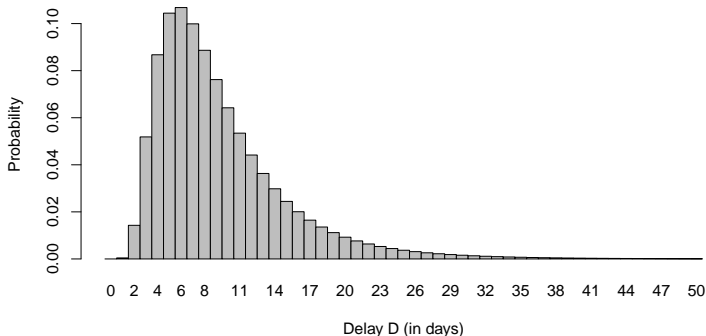
- There is a time delay between time of infection and the onset of the disease. This time delay is often denoted *incubation time*.
- Usually, only onset of disease can be observed. Examples:
 - Time to AIDS onset after HIV infection
 - Onset of diarrhea after consumption of sprouts (STEC/HUS)
- Let D be a discrete random variable describing the delay in number of time units. Assuming this delay is constant over time let $f(d)$, $d = 0, 1, 2, \dots$, be the PMF of D .

Back-projection

Interest is often in the time of exposure of individuals, but data is only available about their time of disease onset.

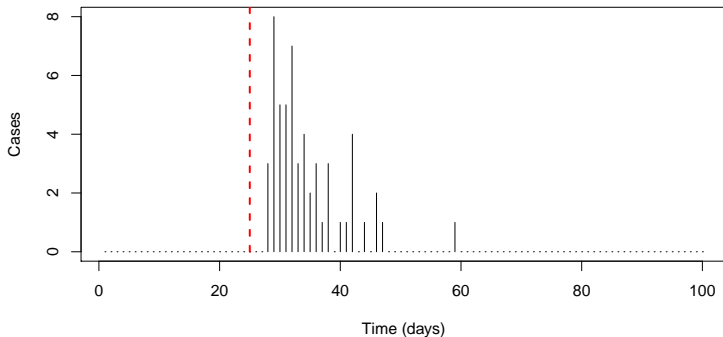
Incubation time as a random variable

- Example: D as discretized version of a log-normal distribution with $\log \mu = 2$, $\log \sigma = 0.6$ and $d_{\max} = 50$.



Example 1: Point source outbreak at time t_0

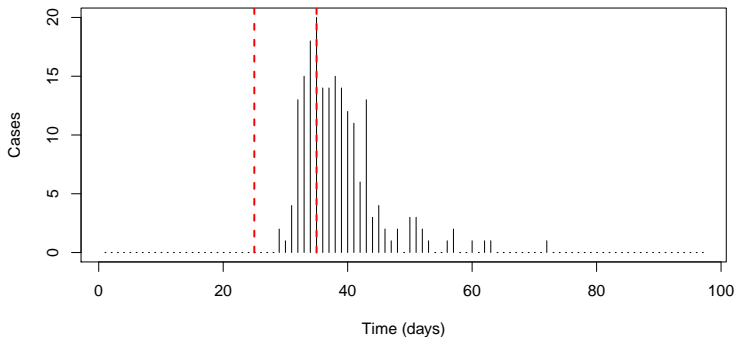
- Assume a point source is active on day $t_0 = 25$ infecting a total of $n = 55$ individuals and f_D as in the previous example.
- The following time series for disease onsets is observed:



- To identify the possible source, interest is in inferring infection times from the onset times.

Example 2: Point source during an interval

- Assume a point source is active for l days from day t_0 on infecting a total of n individuals, where individuals are equally likely to be infected within $[t_0, t_0 + l - 1]$.
- Example $t_0 = 25$, $l = 10$ and $n = 200$.



Simple back-projection methods (1)

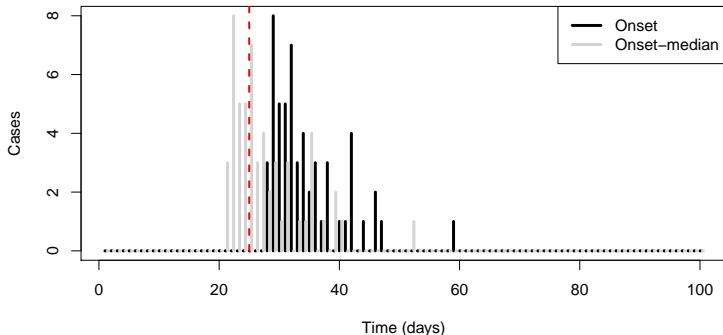
- Method 1: Determine the exposure interval by subtracting the shortest incubation time from the first case and the longest incubation from the last case of the epidemic curve
- R-code for outbreak Examples 1 & 2

```
subtract.minmax <- function(y, d.pmf,eps=1e-3) {
  exposure.left <- head(which(y>eps),n=1) - ((0:d.max)[head(which(d.pmf>eps),n=1)])
  exposure.right <- tail(which(y>eps),n=1) - ((0:d.max)[tail(which(d.pmf>eps),n=1)])
  structure( c(exposure.left,exposure.right-exposure.left),names=c("t0","l"))
}
subtract.minmax(y.ts, d.pmf)
## t0  1
## 26  1
subtract.minmax(y.l.ts, d.pmf)
## t0  1
## 27 13
```

Simple back-projection methods (2)

- Method 2: Subtract the median incubation time from each onset.

```
subtract.median <- function(y,d.pmf) {  
  d.median <- (0:length(d.pmf)-1)[which(cumsum(d.pmf)>0.5)][1]  
  structure(c(tail(y,n=-d.median),rep(0,d.median)),names=names(y))  
}  
subtract.median(y.ts,d.pmf)
```



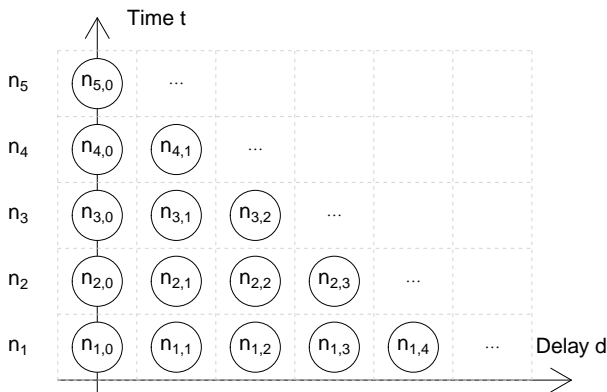
- This method is not recommendable since it ignores the order of events in the epidemic curve.

Non-parametric back-projection by Becker et al.

- Becker et al. proposed a non-parametric back-projection method for discrete time interval data.
- Their motivating application was a back-projection of AIDS cases to HIV incidence (before the use of antiretroviral therapy).
- The method differs from the the individual based continuous time parametric back-calculation of Brookmeyer et al.
- However, it equally presumes a fixed and known incubation time distribution.

Model and notation (1)

$n_{t,d}$ – Number of individuals exposed in interval $t = 1, \dots, T$ having an incubation of time d (i.e. observed at time $t + d$)



y_t – The observed number of incident cases in interval t

$$y_t = \sum_{i=1}^t n_{i,t-i}, \quad t = 1, \dots, T.$$

Model and notation (2)

n_t – Number of individuals infected in interval t , i.e.

$$n_t = \sum_{d=0}^{\infty} n_{t,d}.$$

- Assume $n_t \sim \text{Po}(\lambda_t)$ and as a consequence

$$n_{t,d} \sim \text{Po}(f(d)\lambda_t),$$

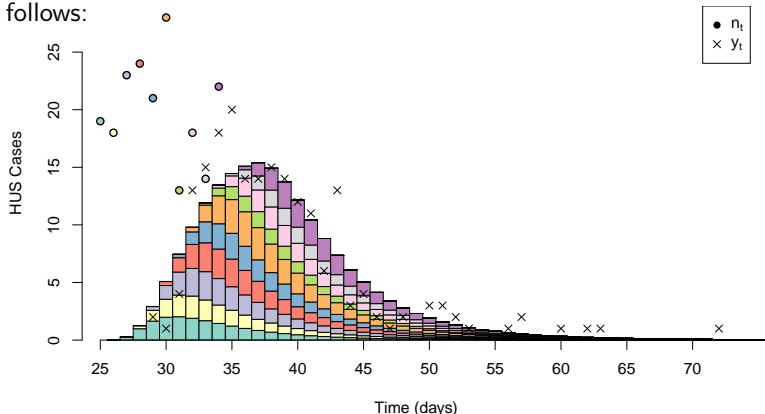
where $f(\cdot)$ is the PMF of the incubation time.

- As a consequence $y_t \sim \text{Po}(\mu_t)$, where

$$\mu_t = \sum_{i=1}^t E(n_{i,t-i}) = \sum_{i=1}^t f(t-i)\lambda_i.$$

Model and notation (3)

- The convoluted μ_t from the previous foil can be illustrated as follows:



- Thus backprojection is the inverse problem of deducing the λ_t 's given the observed y_t 's.

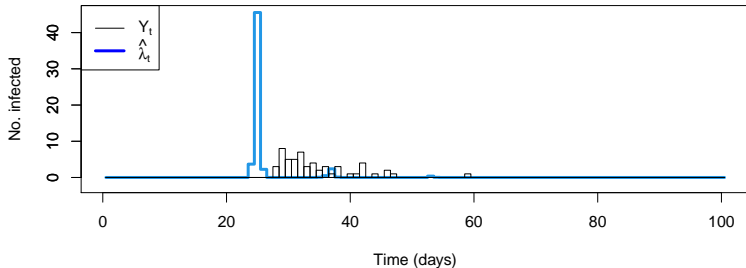
Implementation in surveillance

- Code:

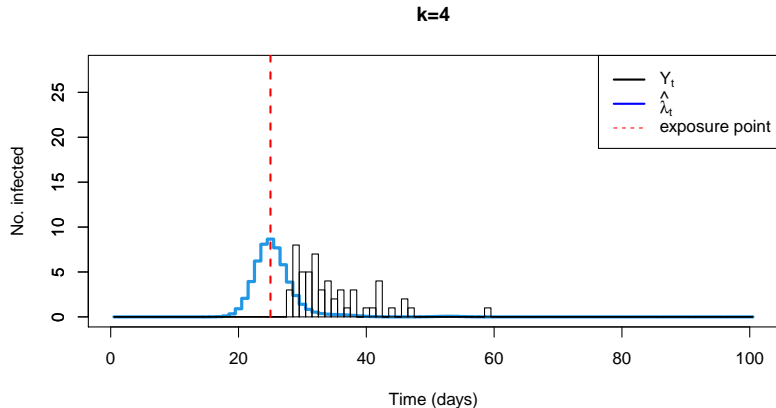
```
#Create vector with incubation time PMF values on (0,...,d_max)
incu.pmf <- c(0, (plnorm(1:d.max,logmu,logsd) - plnorm(0:(d.max-1),logmu,logsd))/plno
#Create sts object
require("surveillance")
sts <- new("sts",epoch=1:length(y.ts),observed=matrix(y.ts,ncol=1))
#Backproject using the method by Becker et al. (1991)
bp.control <- list(k=0,eps=1e-3,iter.max=100,verbose=TRUE,eq3a.method="C")
sts.bp.k0 <- backprojNP(sts, incu.pmf=incu.pmf, control=bp.control)
```

- Plotting code:

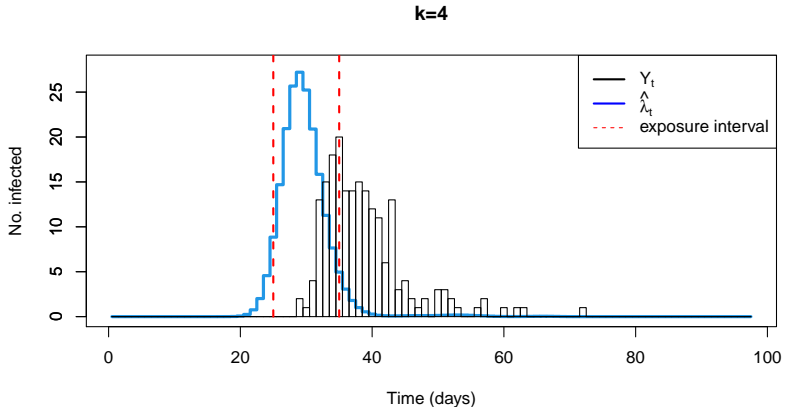
```
plot(sts.bp.k0,xaxis.labelFormat=NULL)
```



Back-projection for outbreak Example 1



Back-projection for outbreak Example 2

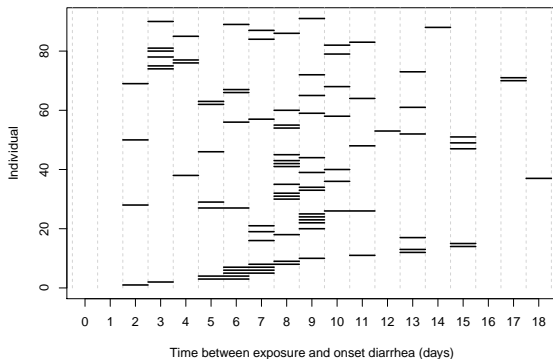


Uncertainty of the estimates

- Problem: The non-parametric back-projection (NPBP) does not provide any measures of uncertainty for the estimate $\hat{\lambda}$
- Two sources of uncertainty exists:
 - Sampling variation in the observed y_t
 - Uncertainty in the estimation of the incubation time

Estimation of the incubation time (1)

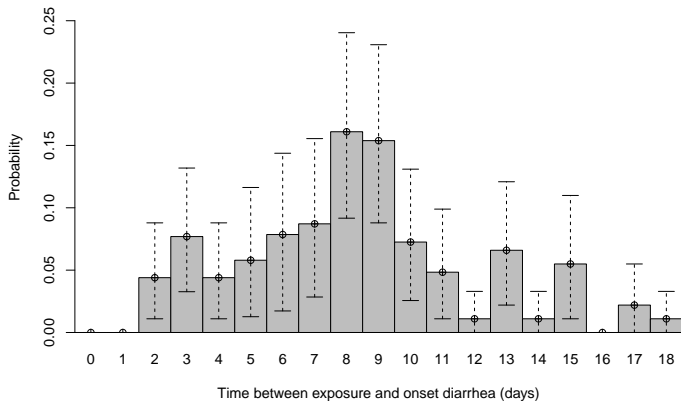
- Determination of the incubation time PMF from 91 cases with a well known exposure time (foreign cases, restaurant cluster, etc.)



- Goal: Non-parametric estimate of the probability mass function

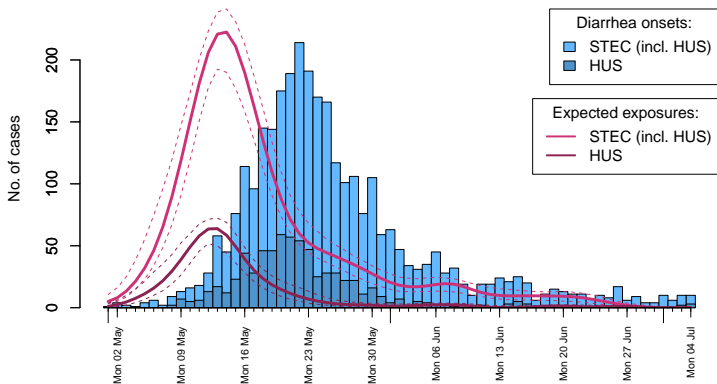
Estimation of the incubation time (2)

- Estimated PMF using Turnbull's method (Turnbull) for interval censored data and point-wise 95% CIs by the percentile method on $R = 999$ additional bootstrap samples



Back-projection for the 2011 STEC/HUS outbreak (4)

- Werber et al. (2013) refines the incubation time estimation by using a Weibull interval censored regression model adjusting for age, sex and HUS in 114 symptomatic adults from six cohorts.



Discussion

- The non-parametric method needs no underlying assumptions about the mode of transmission (person to person, point source, etc.).
- During an outbreak one should choose T such that the incidence cases observed at time y_T are reliable (i.e. sufficiently complete), i.e. T should not be too close to “now”.
- A good recent review of back-projection methods can be found in Egan et al.

Outline

1 Back-calculation method

2 Nowcasting

Nowcasting – what's the situation?

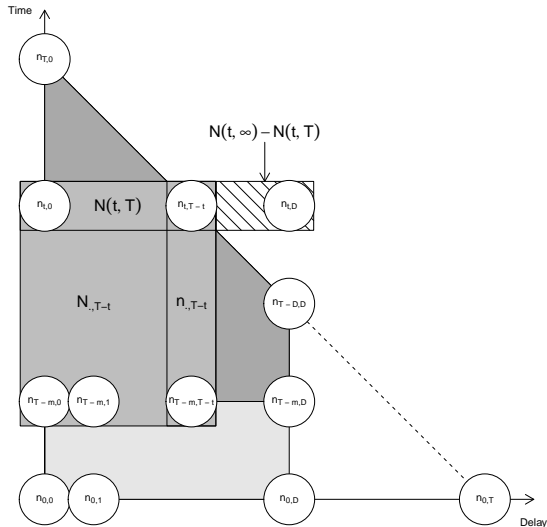
- Opposite to forecasting, we just want to know what the situation is “now” during an outbreak, i.e. in a ideal setup of no reporting delay → nowcasting.
- The term is basically a revival of what has been extensively studied as *reporting delay* during the AIDS/HIV epidemic, see e.g. Kalbfleisch et al. and Harris.
- Nowcasting was used for real-time tracking daily hospitalizations during the 2009 A/H1N1 influenza (Donker et al.).
- There is a close connection between nowcasting and *claims reserving* in actuarial sciences (England et al.).

Nowcasting Notation (1)

- Let $n_{t,d}$ be the number of cases which occur on day t and become available with a delay of d days, where $t = 0, \dots, T$ – with T being *now* – and $d = 0, \dots, D$.
- Problem: $n_{t,d}$ is unknown when $d > T - t$ – see reporting triangle
- $N(t, T) = \sum_{d=0}^{\min(T-t, D)} n_{t,d}$ is the number of cases which occurred on t and who are reported until time T
- Aim of nowcasting: predict the total number of cases, i.e.

$$N(t, \infty) = \sum_{d=0}^{\infty} n_{t,d} = \sum_{d=0}^D n_{t,d}.$$

Nowcasting Notation (2) – Reporting triangle



Nowcasting Methods (1)

- Alternative: The reporting delay for an event follows a distribution with probability mass function $f(d) = f_d$, $d = 0, 1, \dots, D$.
- We will assume time homogeneity of the delay distribution
- Let $F(d) = \sum_{x=0}^d f(x)$ be the CDF of the delay distribution.
- Lawless presents the following nowcast procedure

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

where the CDF F is estimated taking the right-truncation of the data into account, for example by using the reverse time hazard function.

Nowcasting Methods (2)

- Alternative model in Donker et al.

$$N(t, T) \sim \text{Bin} \left(N(t, \infty), \hat{F}(T - t) \right)$$

- In this model inference is about estimating the size parameter in a binomial distribution, i.e.

$$\hat{N}(t, \infty) = \arg \max_{n \geq N(t, T)} \left\{ f_{\text{Bin}}(n, \hat{F}(T - t)) \right\}$$

Nowcasting Methods (3)

- The counts of the reporting triangle can also be thought of as an incomplete contingency table with

$$n_{t,d} \sim \text{Po}(\lambda_t \cdot f_d), \quad t = 0, \dots, T, \quad 0 \leq d \leq \min(T - t, D),$$

where λ_t is the expected number of new events occurring at time t .

- Altogether, $T + D + 2$ parameters are to be estimated.
- The above presentation lends itself to log-linear modeling, i.e. with parametric, semi-parametric or non-parametric linear predictor

$$\log \mu_{t,d} = \log(\lambda_t) + \log(f_d) = s(t; \beta) + v(d; \theta),$$

where $E(n_{t,d}) = \mu_{t,d}$.

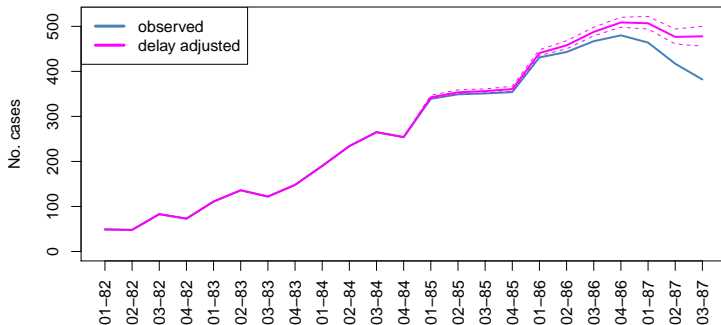
Example: AIDS registry data from the CDC (1)

- Zeger et al. contains an analysis of 6190 homosexual AIDS cases classified by quarter of diagnosis and the number of quarters between diagnosis and report to the CDC

##	0	1	2	3	4	5	6	7	8	9	10	11	12
## 03-87	244	138	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 02-87	217	165	35	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 01-87	317	80	54	13	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 04-86	353	64	32	17	14	NA	NA	NA	NA	NA	NA	NA	NA
## 03-86	345	53	35	18	10	6	NA	NA	NA	NA	NA	NA	NA
## 02-86	313	60	29	15	10	9	7	NA	NA	NA	NA	NA	NA
## 01-86	294	71	27	10	13	5	5	6	NA	NA	NA	NA	NA
## 04-85	216	68	23	21	10	5	2	6	3	NA	NA	NA	NA
## 03-85	206	58	36	23	9	7	2	6	0	4	NA	NA	NA
## 02-85	215	61	26	19	15	3	4	4	1	0	1	NA	NA
## 01-85	188	70	36	22	11	6	2	3	1	0	0	0	NA
## 04-84	159	36	20	14	9	1	4	2	2	1	1	3	2
## 03-84	149	51	26	16	10	6	3	0	1	0	0	0	3
## 02-84	140	51	16	12	3	5	1	3	1	1	0	1	0
## 01-84	119	41	9	4	8	2	3	2	0	1	0	0	1
## 04-83	94	26	9	8	4	0	2	0	0	1	2	0	2
## 03-83	80	24	5	3	3	2	1	1	0	0	0	1	2
## 02-83	97	23	9	0	1	1	1	2	0	1	0	1	0
## 01-83	67	25	7	7	1	1	0	1	0	2	0	0	0
## 04-82	52	10	6	1	2	0	1	0	0	0	0	0	1
## 03-82	59	11	7	1	2	1	1	0	0	0	0	0	1
## 02-82	36	4	1	0	3	1	1	0	1	0	0	0	1
## 01-82	24	8	5	0	4	2	1	0	1	0	0	0	1

Example: AIDS registry data from the CDC (2)

- The model is an instance of a generalized linear model, which can be fitted in R using the function `glm`
- Point estimate for the delay adjusted $N(t, \infty)$'s in the AIDS example (+ pointwise predictive distributions).



Discussion

- In practice the delay distribution often time-inhomogeneous. In this situation a proportional hazards model for the reverse time hazard function can be used (Kalbfleisch et al.; Pagano et al.).
- Back-projection based on registry data for an ongoing epidemic is often to be seen concurrently with delay adjustments (Brookmeyer et al.; Zeger et al.; Kalbfleisch et al.).

Literature I



Becker, N. G., L. F. Watson, and J. B. Carlin. "A method of non-parametric back-projection and its application to AIDS data". Statistics in Medicine 10 (1991): 1527–1542.



Brookmeyer, R., and A. Damiano. "Statistical methods for short-term projections of AIDS incidence". Stat Med 8, no. 1 (Jan. 1989): 23–34.



Brookmeyer, R., and M. Gail. "A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic". Journal of the American Statistical Association 83 (1988): 301–308.



Donker, T., M. van Boven, W. M. van Ballegooijen, T. M. Van't Klooster, C. C. Wielders, and J. Wallinga. "Nowcasting pandemic influenza A/H1N1 2009 hospitalizations in the Netherlands". European Journal of Epidemiology 26, no. 3 (2011): 195–201.

Literature II



Egan, J. R., and I. M. Hall. “A review of back-calculation techniques and their potential to inform mitigation strategies with application to non-transmissible acute infectious diseases”. J R Soc Interface 12, no. 106 (May 2015).



England, P., and R Verrall. “Stochastic Claims Reserving in General Insurance, Institute of Actuaries and Faculty of Actuaries”. Online at <http://www.actuaries.org.uk/sessional/sm0201-report.html>, Faculty and Institute of Actuaries, Sessional Meeting Paper (2002).



Harris, J. E. “Reporting Delays and the Incidence of AIDS”. JASA 85, no. 412 (1990): 915–924.



Kalbfleisch, J. D., and J. F. Lawless. “Inference Based on Retrospective Ascertainment: An Analysis of the Data on Tranfusion Related AIDS”. Journal of the American Statistical Association 84, no. 406 (1989): 360–372.

Literature III



– . “Regression models for right truncated data with applications to AIDS incubation times and reporting lags”. Statistica Sinica 1 (1991): 19–32.



Lawless, J. F. “Adjustments for Reporting Delays and the Prediction of Occurred but Not Reported Events”.
The Canadian Journal of Statistics 22, no. 1 (1994): 15–31.



Pagano, M., X. M. Tu, V. De Gruttola, and S. MaWhinney. “Regression analysis of censored and truncated data: estimating reporting-delay distributions and AIDS incidence from surveillance data.” Biometrics 50, no. 4 (1994): 1203–1214.



Turnbull, Bruce W. “The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data”.
Journal of the Royal Statistical Society. Series B (Methodological) 38, no. 3 (1976): pp. 290–295. ISSN: 00359246.
<http://www.jstor.org/stable/2984980>.

Literature IV



Zeger, S. L., L. C. See, and P. J. Diggle. “Statistical methods for monitoring the AIDS epidemic”. Stat Med 8, no. 1 (Jan. 1989): 3–21.