Inferring Reproductive Numbers from Final Size Epidemic Data on Hepatitis A in Flanders, Belgium: Truncation, Censoring and Heterogeneity

Niel Hens

InFER - Warwick 2011

Outline

Introduction

- Case Study on Hepatitis A
- 'Federal' knowledge centre:

  If any, targeted or universal vaccination?
- Five steps:
  - Immunological assessment
  - Serological assessment
  - Outbreak modelling
  - Mathematical modelling
  - Cost-effectiveness analysis

Hepatitis A Immunology

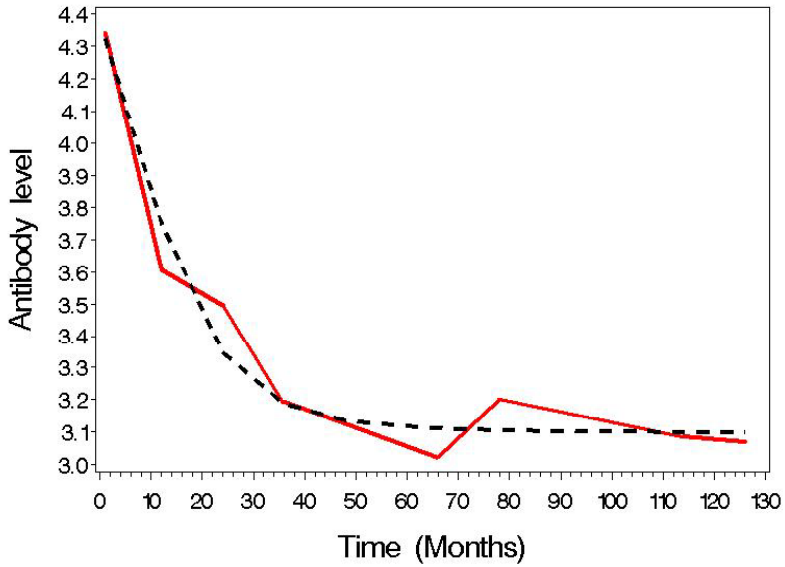
- Hepatitis A vaccine study: post-vaccination antibody levels follow-up over 10 years (113 individuals)
- Consider long and short living plasma cells producing antibodies

$$\begin{aligned} \frac{dP_s}{dt} &= -\mu_s P_s, \\ \frac{dP_l}{dt} &= -\mu_l P_l, \\ \frac{dA}{dt} &= \phi_s P_s + \phi_l P_l - \mu A, \end{aligned}$$

  where $\mu_.$ are mortality rates and $\phi_.$ are production rates.
- Analytical solution
- Inference using nonlinear mixed models (SAS & Monolix)

Hepatitis A Immunology

- Results:
  - antibody levels: lifespan of 1 month
  - short-lived plasma cells: average lifespan of 7 months
  - long-lived plasma cells persist for life



- Assumption of lifelong immunity seems reasonable

Hepatitis A Serial Seroprevalence

- Serological samples from 1993 and 2002
- Serial seroprevalence
- Ades and Nokes [1993], Batter et al. [1994], Marschner [1997], Nagelkerke et al. [1999]: multiplicative age & time specific model

$$\lambda(a,t) = \exp(\mu_0 + \sum_{i=1}^{I} \mu_i a^i + \sum_{j=1}^{J} \theta_j t^j), \quad I, J = 1, 2, ...,$$

  assuming differential selection and a (semi-)parametric model.
- Focus:
  - hepatitis A: SIR
  - non-differential selection
  - general framework

Hepatitis A Serial Seroprevalence

- Nagelkerke et al. [1999]: proportional hazards model
- for two calendar times t_1 and t_2:

$$\lambda(a, t_2) = \exp(\beta(t_2 - t_1))\lambda(a, t_1). \tag{1}$$

- a cohort born at calendar time b has age a at calendar time $a + b$:

$$\lambda_b(a) = \lambda(a, a + b).$$

- the proportional hazards assumption:

$$\lambda_b(a) = \exp(\beta(b - b_0))\lambda_{b_0}(a), \tag{2}$$

  where $\lambda_{b_0}(a) = \lambda(a, a + b_0)$ is the baseline force of infection
- MM-algorithm: maximization-maximization with isotonic stepwise estimate to achieve $\lambda_{b_0}(a) \geq 0$

Hepatitis A Serial Seroprevalence

- fully nonparametric proportional hazards model:

$$\lambda(a, t_2) = \exp(g(t_2 - t_1))\lambda(a, t_1) \tag{3}$$

where $g(t)$ is a smooth function with the constraint that $g(0) = 0$.

- In terms of the proportion susceptible

$$x_b(a) = x(a, b + a) = \exp\{\exp(g(b - b_0)) \log x_{b_0}(a)\}, \tag{4}$$

where, with $x_{b_0}(a) = x(a, b_0 + a)$,

$$x_{b_0}(a) = \exp\{-\int_0^a \lambda_{b_0}(s)ds\}.$$

Hepatitis A Serial Seroprevalence

- Denote $\pi_b(a) = 1 - x_b(a)$ the proportion subjects infected at age a or before:

$$\log(-\log(1 - \pi_b(a))) = s_1(a) + s_2(b), \tag{5}$$

where

  - $s_1(a) = \log(-\log(1 - \pi_{b_0}(a)))$.
  - $s_2(b) = g(b - b_0)$,

- Cohort-specific force of infection:

$$\lambda_b(a) = \frac{\pi_b'(a)}{1 - \pi_b(a)} = \exp(s_2(b))\lambda_{b_0}(a), \tag{6}$$

with

$$\lambda_{b_0}(a) = \frac{\pi_{b_0}'(a)}{1 - \pi_{b_0}(a)} = \frac{d}{da}\exp(s_1(a)). \tag{7}$$

- Extensions are possible

Hepatitis A Serial Seroprevalence

- Standard statistical software: 'mgcv'-package in R: thin plate regression splines
- Test for cohort-effect: approximate F-test
- Results:

	Model Components	edf	AIC	rank	BIC	rank
Model 1:	$a + b$	3	6062.03	5	6082.07	4
Model 2:	$s(a) + b$	10.03	5976.90	4	6043.94	2
Model 3:	$a + s(b)$	10.55	5958.68	3	6082.07	5
Model 4:	$s(a) + s(b)$	13.85	5934.30	1	6026.83	1
Model 5:	$s(a, b)$	19.04	5936.16	2	6063.34	3

Table: Results for different models: parametric model 1, semi-parametric models 2 and 3, non-parametric models 4 and 5.

- Note: TB example of Nagelkerke et al. [1999]: Model 5 - best model

Hepatitis A Serial Seroprevalence

- Ensuring positivity for the cohort-specific FOI:
  - maximization-maximization
  - age: monotone P-spline
  - time: P-spline
- Backfitting algorithm
- Iterate until convergence
- Two-dimensional tuning parameter selection

Hepatitis A Serial Seroprevalence

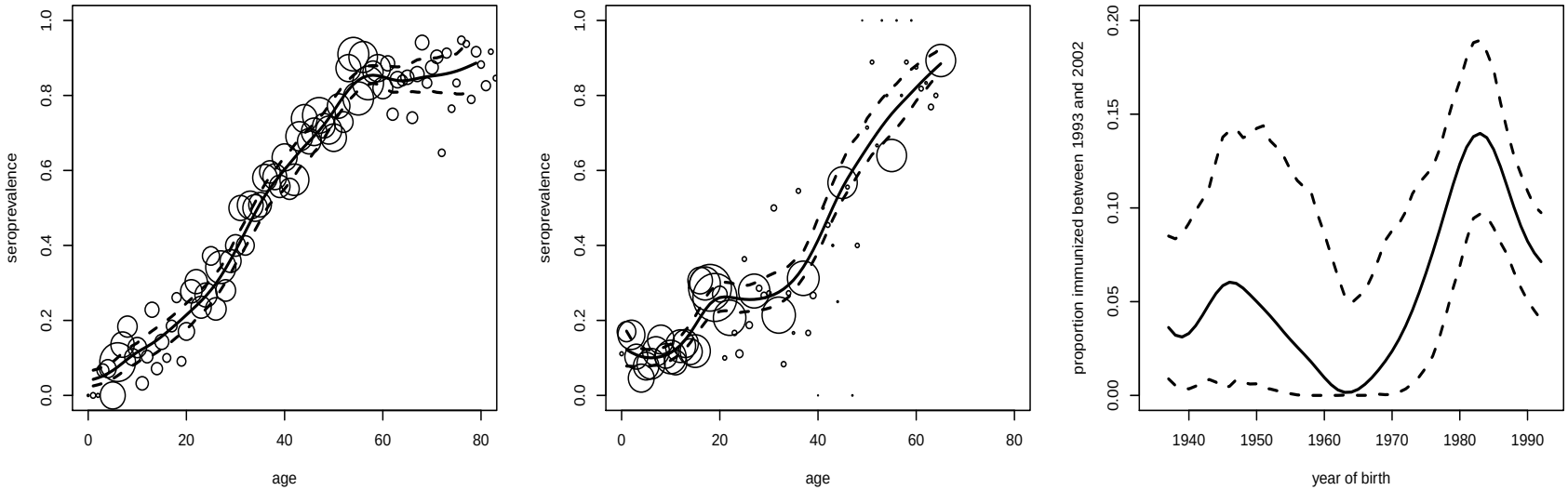
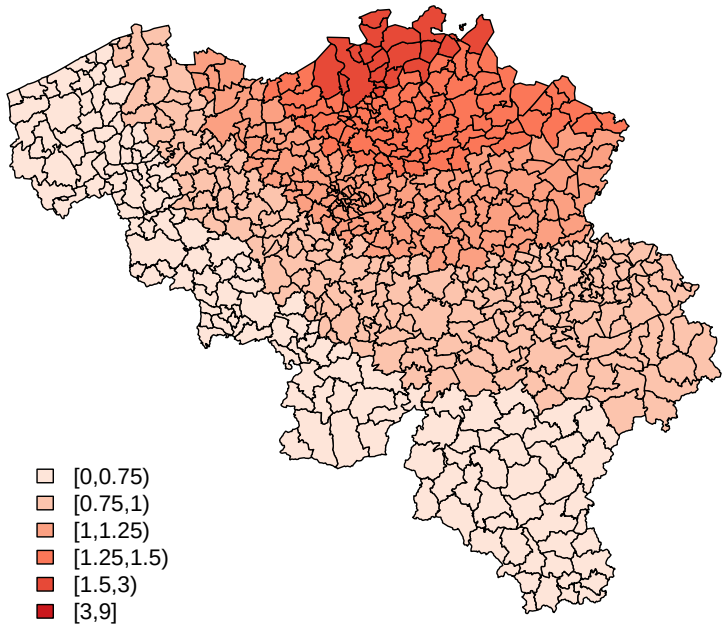


Figure: HAV-seroprevalence in Flanders based on serological sample from 1993 (left panel) and 2002 (middle panel) together with the proportion infected in the period 1993-2002 (right panel). Model-based estimates (solid lines) and 95% bootstrap-based percentile confidence intervals (dashed lines) are based on the analysis reported by Hens et al. [2011].

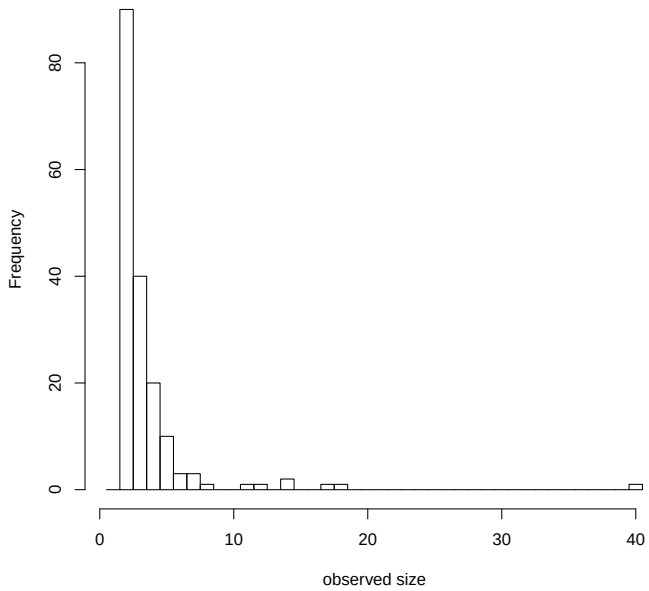
Hepatitis A Serial Seroprevalence

- Seroprevalence data 2002 for Belgium
- Odds ratios for the spatial effect (tensor-product spline for age)



Hepatitis A Outbreak Data

- 113 outbreaks (provincial health agencies)
- 5 foodborne outbreaks: excluded from the analysis
- available information:
  - final sizes
  - 5 Flemish provinces
  - auxiliary information: school, family, children, ...
- reporting issues:
  - number of initial cases mostly unknown
  - outbreaks with size at least 2
  - underreporting due to long incubation period: 15-45 days



Objectives

- Estimate the reproduction number from final size data
- Dealing with reporting issues:
  - marginalization
  - truncation
  - censoring
  - heterogeneity
- Input for a mathematical model with
  - demography
  - migration rates

  $\rightarrow$ confirmation of no transmission, only importation

Estimating the Reproduction Number

- Mortal branching process
- Assume we know the number of initial cases
- Final size: total number of infected persons with links
- Becker (1974): final size distribution

$$P(X = x; s) = b(x, s) \frac{\theta^{x-s}}{A(\theta)^x},$$

where

- $S = s$ initial cases
- $X = x$ is the final size
- θ is the reproductive ratio
- $b(x, s)$ is constant.
- $A(\theta)^x$ is a normalizing factor

Estimating the Reproduction Number

- Offspring distribution: generalized power series distribution
- Poisson offspring distribution: Borel-Tanner distribution [Haight and Breuer, 1960]

$$P(X = x; s) = \frac{s x^{x-s-1} \theta^{x-s} e^{-x\theta}}{(x-s)!}, \quad x = s, s+1, \ldots.$$

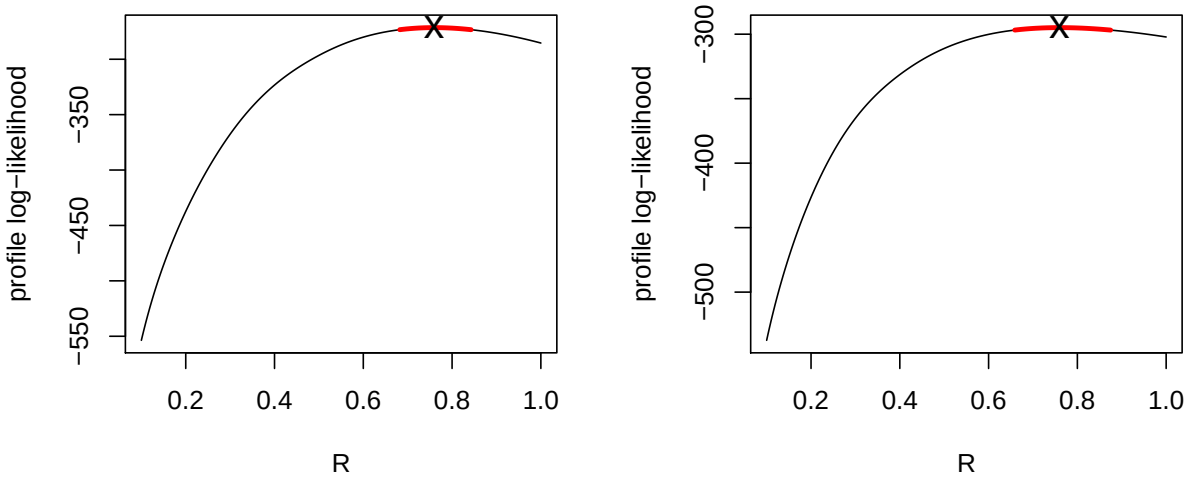
- Geometric offspring distribution:

$$P(X = x; s) = \frac{s}{2x-s} \begin{pmatrix} 2x-s \\ x-s \end{pmatrix} \frac{\theta^{x-s}}{(1+\theta)^{2x-s}}, \quad x = s, s+1, \ldots.$$

- Defined for $X < \infty$ and $\theta \leq 1$

Applied to the Data

- Assume 1 source case for each outbreak (ignoring truncation):



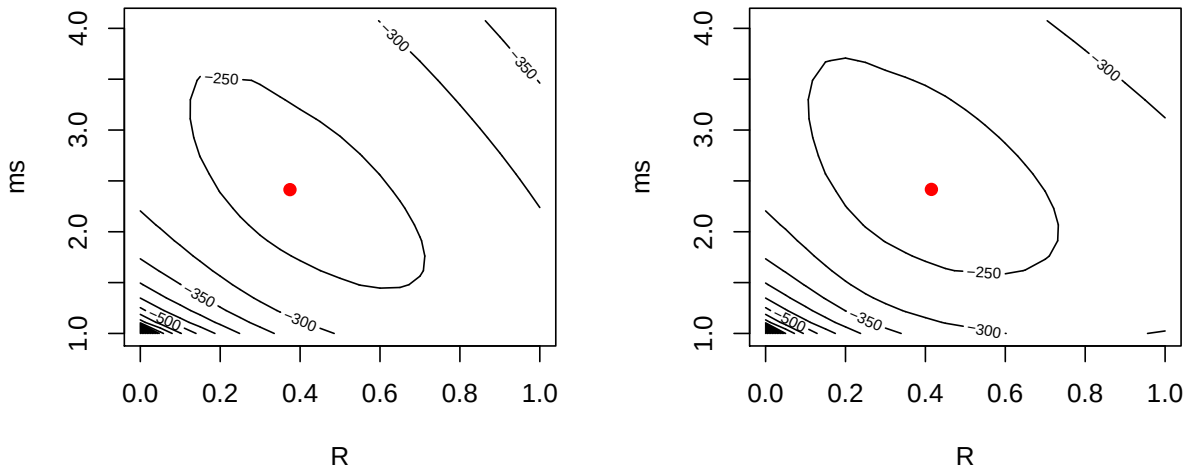
- profile likelihood confidence intervals
- Borel-Tanner: $\hat{R} = 0.76\ (0.68, 0.84)$
- Geometric Offspring: $\hat{R} = 0.76\ (0.66; 0.88)$

Unknown Number of Initial Cases

- Crucial assumption: conditioning $S = s$
  - assume initial case distribution
  - derive the joint and marginal likelihood $\{x \geq s\}$
- Initial case distribution
  - one-parameter distributions
    - degenerate: $S = 1$
    - discrete cases: $S = 1, 2$
    - truncated Poisson
  - two-parameter distributions
    - discrete cases: $S = 1, 2, 3$
    - truncated negative binomial

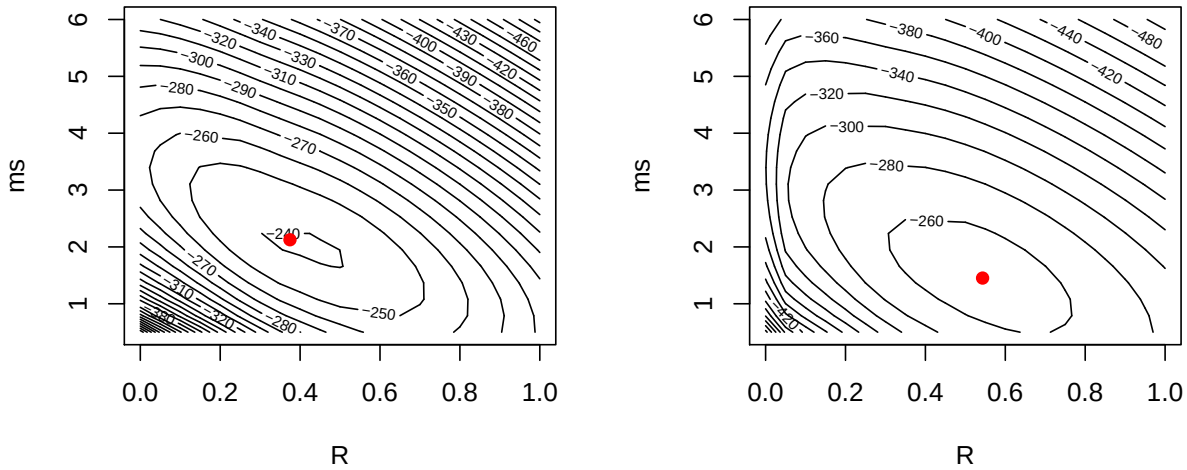
Applied to the Data

- Unknown initial cases (ignoring truncation):
- Illustration: truncated Poisson
- Borel-Tanner: $\hat{R} = 0.37$, $\widehat{E(S)} = 2.41$
- Geometric offspring: $\hat{R} = 0.42$, $\widehat{E(S)} = 2.41$



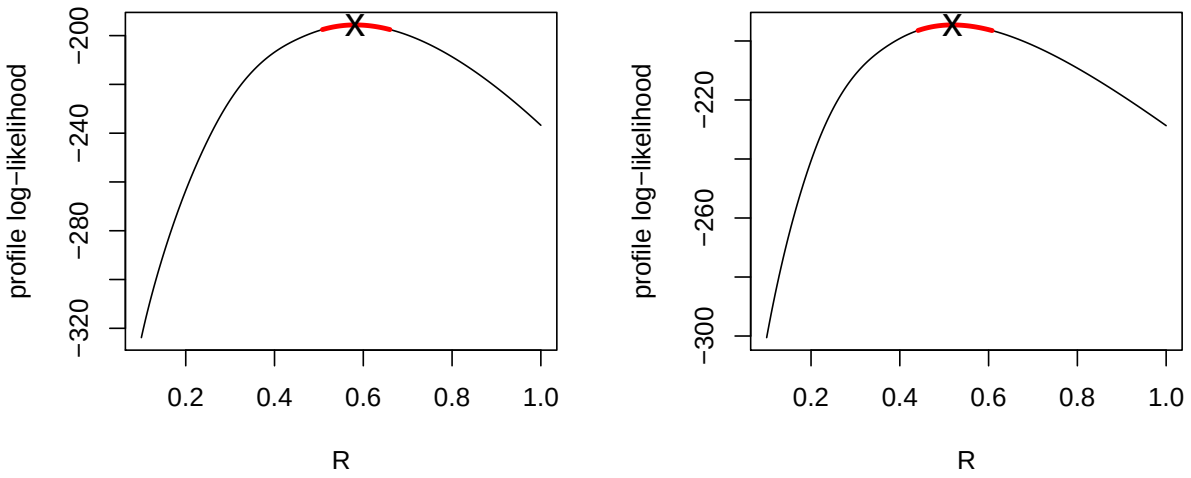
Known and Unknown Number of Initial Cases

- Marginalization induces assumption dependent results
- What if we do have some information on initial cases (7×1)?
  - ignorance leads to inefficiency provided a correct model is specified
  - missing data approach? direct likelihood
- Illustration: truncated Poisson & Borel-Tanner



Truncation: Known Number of Initial Cases

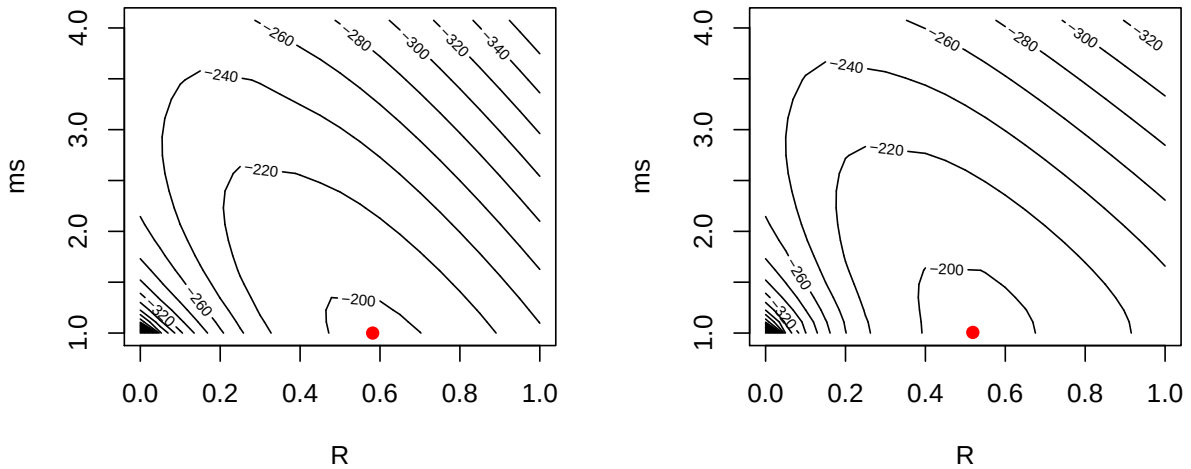
- Assume 1 source case for each outbreak:



- profile likelihood confidence intervals
- Truncated Borel-Tanner: $\hat{R} = 0.58\ (0.51, 0.66)$
- Truncated Geometric Offspring: $\hat{R} = 0.52\ (0.44; 0.61)$

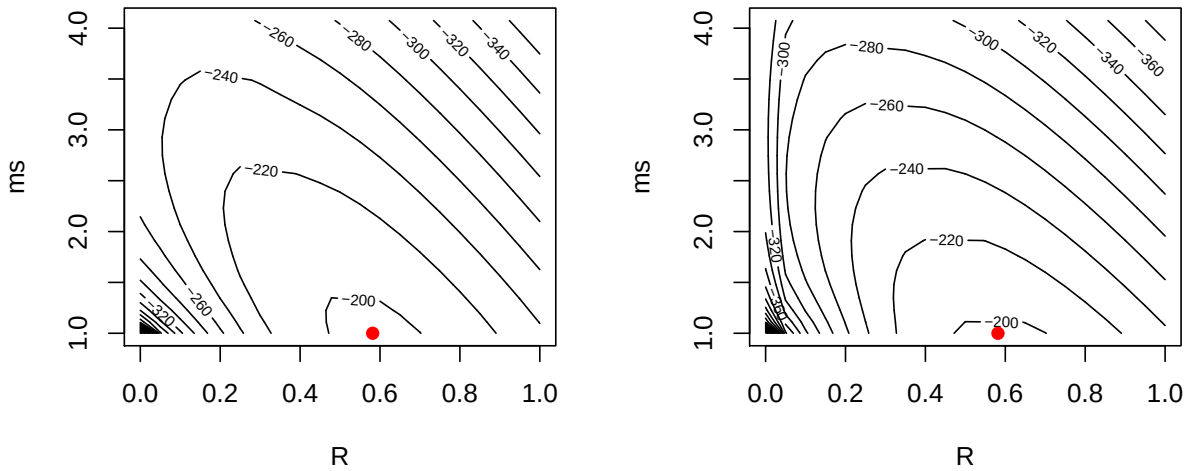
Truncation: Unknown Number of Initial Cases

- Illustration: truncated Poisson
- Truncated Borel-Tanner: $\hat{R} = 0.58$, $\widehat{E(S)} = 1.00$
- Truncated Geometric offspring: $\hat{R} = 0.52$, $\widehat{E(S)} = 1.01$



Truncation: Known and Unknown Number of Initial Cases

- Illustration: truncated Poisson
- Truncated Borel-Tanner: $\hat{R} = 0.58$, $\widehat{E(S)} = 1.00$



Censoring & Conditioning on Extinction

- Farrington et al. [2003]:
- $X = \infty$:

$$P(X = \infty) = 1 - q(\theta)^s$$

- Censored likelihood:

$$L(\theta; s, x, c) = \prod_{i=1}^{n} \left\{ P(X = x_i; s_i)^{c_i} \left(1 - \sum_{j=s_i}^{x_i - 1} P(X = j, s_i) \right)^{1-c_i} \right\}$$

- Only two clusters affected by censoring: $\hat{R} = 0.59, \widehat{E(S)} = 1.00$
- Assumption: non-informative censoring

Heterogeneity

- Spread depends on the setting
  - calendar time
  - school vs non-school
  - ...
- Auxiliary information can be used:
  - regression approach:
    - additive effect
    - multiplicative effect
  - random effects approach: ωR where $\omega \sim \Gamma(\theta, 1/\theta)$
    - empirical Bayes estimates for individual effects

Applied to the Data

- Illustration: regression approach & marginal truncated Borel-Tanner
- Provinces (significant effect due to Antwerp):
  - Antwerp: $\hat{R} = 0.84$
  - East-Flanders: $\hat{R} = 0.43$
  - Flemish-Brabant: $\hat{R} = 0.45$
  - Limburg: $\hat{R} = 0.56$
  - West-Flanders: $\hat{R} = 0.57$
- School vs Non-school:
  - School: $\hat{R} = 0.47$
  - Non-school: $\hat{R} = 0.72$

Epilogue

- Results were used as input parameters for the mathematical model
- We used WAIFW and contact data
- Mathematical model with demography, including migration

  $\rightarrow$ confirms no transmission only migration
- Cost-effectiveness analysis

  $\rightarrow$ vaccination of second generation children
- Difficult to 'sell' that story

Acknowledgements

Co-authors: Christel Faes, Joke Bilcke, Steven Abrams, Stijn Jaspers, Yannick Vandendijck, Marc Aerts and Philippe Beutels

Prologue I: Olivier Lejeune, Mathieu Andraud, Zebedee Musoro

Epilogue: Chellafe Ensoy

This work has been funded by "SIMID", a strategic basic research project funded by the institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), project number 060081.

References

A. E. Ades and D. J. Nokes. Modeling age- and time-specific incidence from seroprevalence:toxoplasmosis. *Am J Epidemiol*, 137(9): 1022–1034, May 1993.

V. Batter, B. Matela, M. Nsuami, M. T., M. Kamenga, F. Behets, R. Ryder, W. Heyward, J. Karon, and M. St Louis. High HIV-1 incidence in young women masked by stable overall seroprevalence in young childbearing women in Kinshasa, Zaire: Estimating incidence from seroprevalence data. *AIDS*, 8:811–817, 1994.

C. P. Farrington, M. N. Kanaan, and N. J. Gay. Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, 4(2):279–295, Apr 2003. doi: 10.1093/biostatistics/4.2.279.

F. Haight and M. Breuer. The borel-tanner distribution. *Biometrika*, 47:143–150, 1960.

N. Hens, Z. Shkedy, M. Aerts, C. Faes, P. Van Damme, and P. Beutels. *Infectious Disease Parameters for Transmission Models: A Modern Statistical Perspective*. Springer-Verlag: Forthcoming, 2011.

I. Marschner. A method for assessing age-time disease incidence using serial prevalence data. *Biometrics*, 53:1384–1398, 1997.

N. Nagelkerke, S. Heisterkamp, M. Borgdorff, J. Broekmans, and H. Van Houwelingen. Semi-parametric estimation of age-time specific infection incidence from serial prevalence data. *Statistics in Medicine*, 18:307–320, 1999.