Statistical modelling in public health: segmented regression and methods to model low count data

Dr. School seminar, Summer 2017

Lukas Richter

02.06.2017

Background

- Streptococcus pneumoniae-related infections: among children main cause of
  - meningitis
  - bacterial pneumonia
  - sepsis
- 90 distinct pneumococcal serotypes
- Only a small number account for invasive pneumococcal disease (IPD)
- January 2012: pneumococcal conjugate vaccine introduced in the national childhood immunisation program:
  - covering 10 serotypes (PCV10)
  - administered at 3rd, 5th and 12th month of life
  - funded
- Other vaccines: PCV13, PPV23

Objective

Objective

Our aim was to evaluate direct and indirect longitudinal effects of introducing a 10-valent pneumococcal conjugate vaccine in the Austrian childhood immunization program in 2012 on the occurrence of invasive pneumococcal disease in the Austrian population.

- Analyse case-based data of confirmed IPD from 2009–2016/05
- Use information on age, date of diagnosis and serotype

Datasources

Merged data from 4 sources:

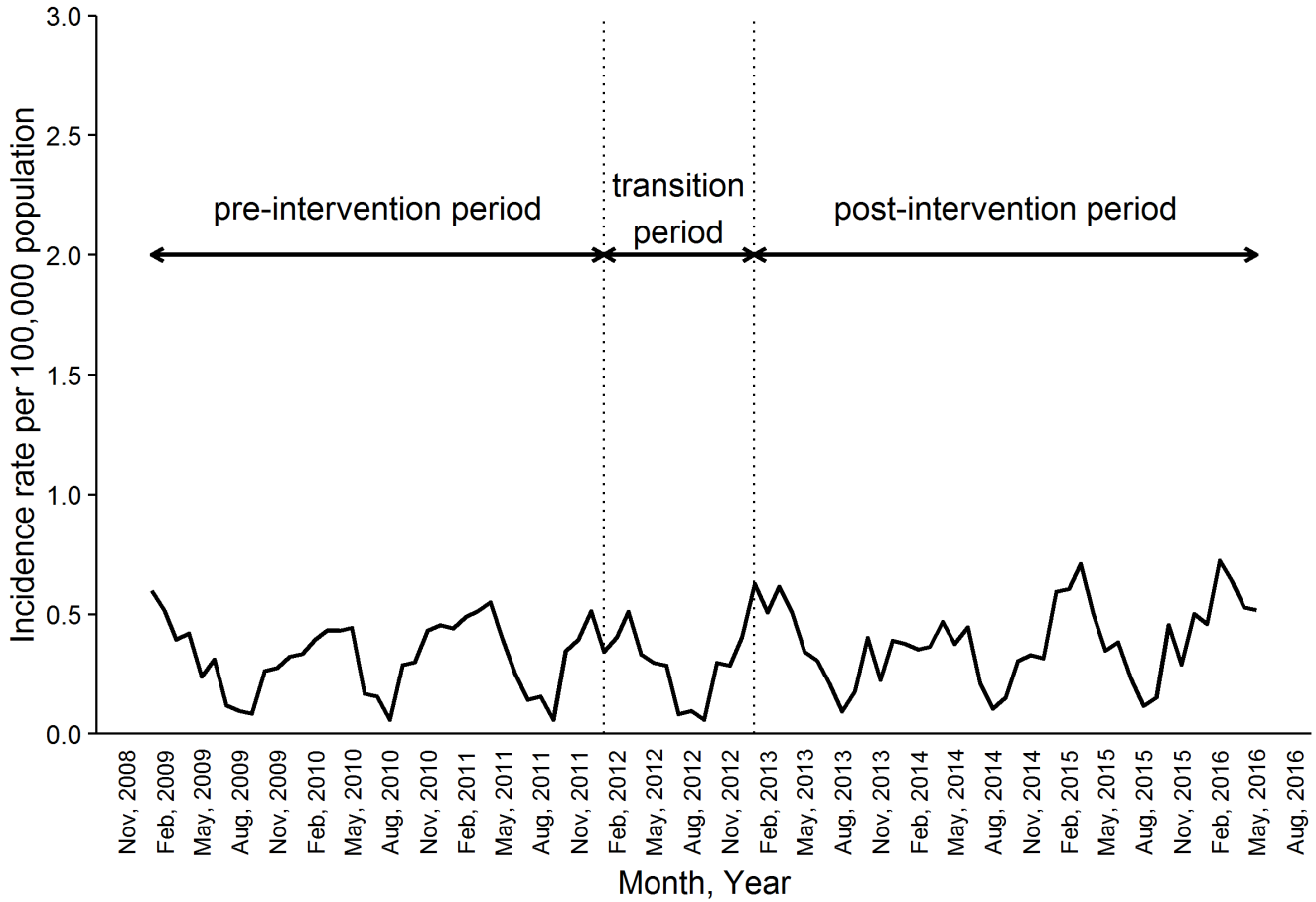
- National surveillance data on IPD, AGES Vienna
- National reference laboratory for IPD, AGES, Graz
- Surveillance-study on IPD among <5 yeras old operated by the Institute of Specific Prophylaxis and Tropical Medicine at the Medical University Vienna
- Population data from Statistics Austria

Method

- Transform data into time series
- Interpret intervention as interruption of the time series
  $\Rightarrow$ interrupted time series
- Segmented regression allows detection of changes when comparing adjacent segments regarding
  - the secular trend
  - one-time (sudden) changes
  - seasonal pattern
- Apply a segmented negative binomial regression model
  - monthly IPD incidence data
  - period: January 2009 – May 2016
  - non-vaccine target groups, 5–49 and $\geq$50 years old
  - serotype subgroups (e.g. PCV10 serotypes)

Study periods

Figure: Monthly incidence rate of total IPD per 100,000 population including model segments (arrows), Austria, 2009 – 2016/05.



Definition of outcome

Table: Definition of the intervention and control outcome.

Age-group	Type of Outcome	Outcome definition
<5	Intervention	PCV10 ST-IPD
	Control	Non PCV10 ST-IPD
	Control	3, 6A, 19A ST-IPD
≥50	Intervention	PCV10 ST-IPD
	Control	Non PCV10 ST-IPD
	Control	PPV23 ST-IPD
	Control	Non PPV23 ST-IPD
	Control	3, 6A, 19A ST-IPD
5–49	Control	PCV10 ST-IPD
	Control	Non PCV10 ST-IPD

Model

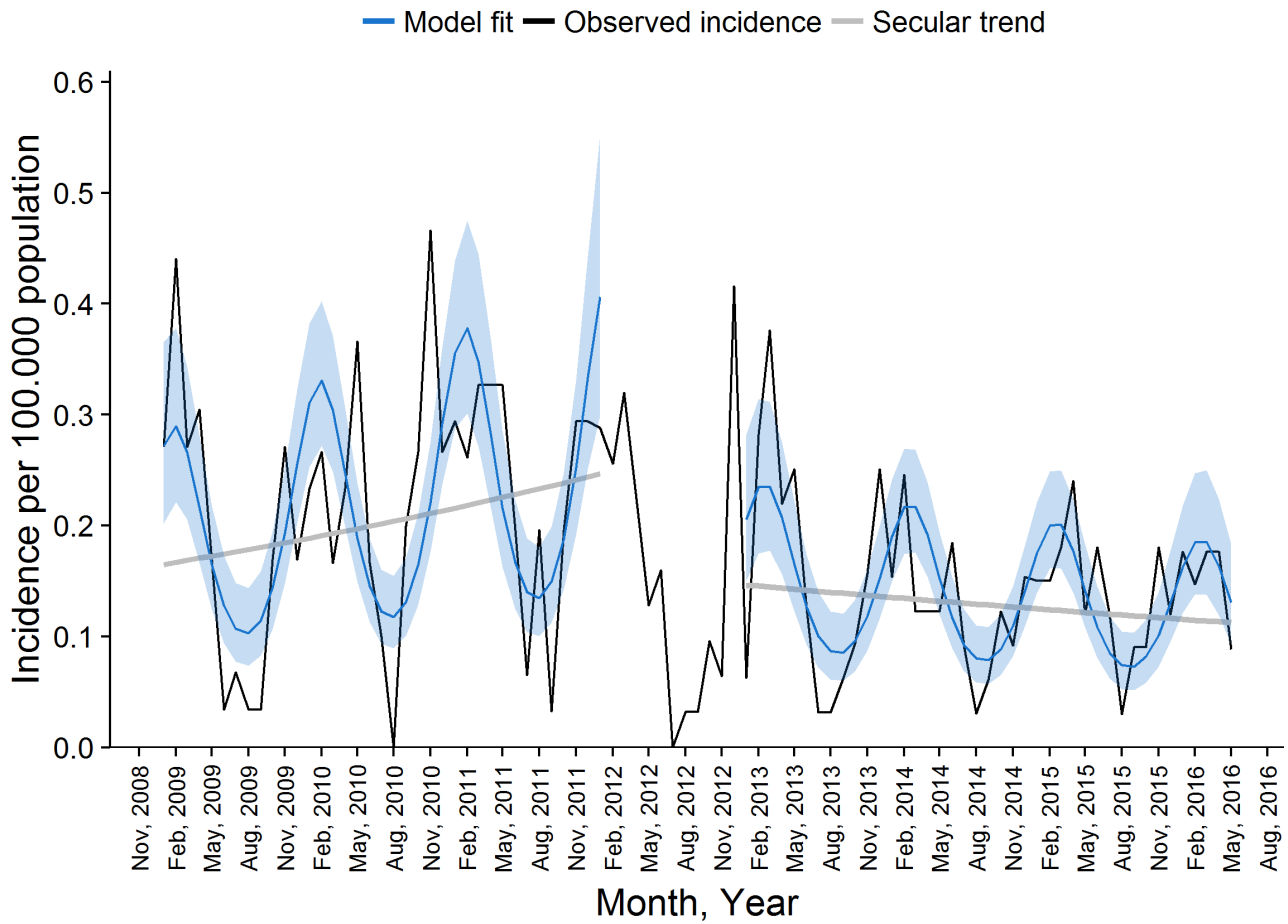
$$\begin{aligned}\log(Y_t) &= \log(\text{pop}_t) + \beta_0 + \beta_1 \,\text{time}_t + \beta_2 \,\sin\left(\frac{2\pi t}{12}\right) + \beta_3 \,\cos\left(\frac{2\pi t}{12}\right) \\ &+ \beta_4 \,\text{vaccine}_t + \beta_5 \,\text{time\_after\_vaccine}_t + \beta_6 \,\sin_{\text{aftervaccine}}\left(\frac{2\pi t}{12}\right) \\ &+ \beta_7 \,\cos_{\text{aftervaccine}}\left(\frac{2\pi t}{12}\right) + \varepsilon_t.\end{aligned}$$

Model Variables

- Y_t: number of IPD cases observed in month t
- pop_t: population
- time_t: time in months
- vaccine_t: binary, indicates if month t is assigned to the pre- ($\text{vaccine}_t = 0$) or to the post-intervention period ($\text{vaccine}_t = 1$)
  $\Rightarrow \text{vaccine}_t = 1$ for $t > 48$ and zero otherwise
- $\text{time\_after\_vaccine}_t$: number of months elapsed in the post-intervention period
  $\Rightarrow \text{time\_after\_vaccine}_t = \text{time}_t - 48$ for $t > 48$ and zero otherwise
- sin, cos, $\text{sin}_{\text{aftervaccine}}$ and $\text{cos}_{\text{aftervaccine}}$: seasonality in pre- and post-intervetion period

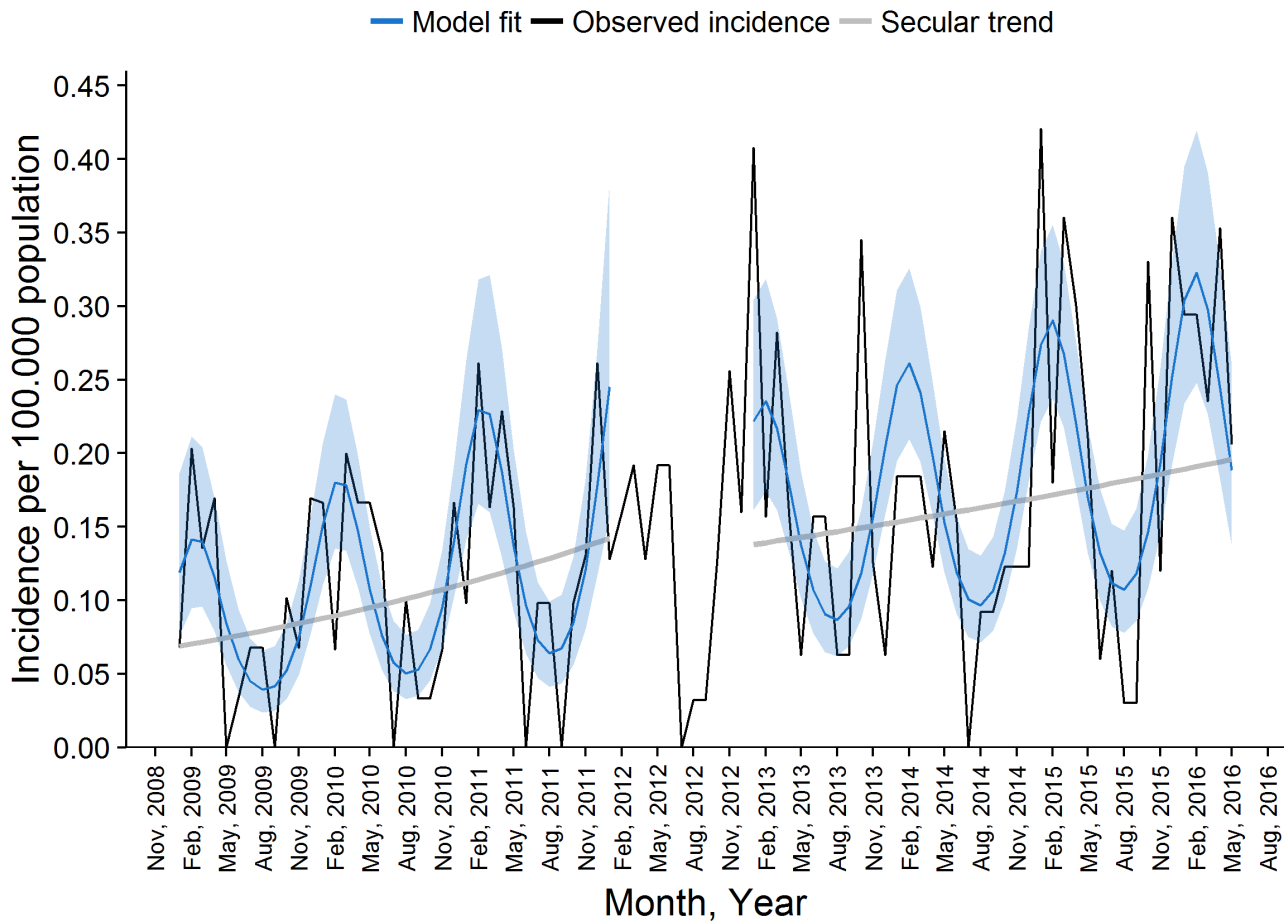
Results: Intervention outcome among ≥50 years old

Figure: Observed incidence/100,000 population, model fit with 95% CI and secular trend of the segmented regression model including 2012 as transition period. *PCV10 ST-IPD* among ≥50 years old, Austria, 2009 – 2016/05.



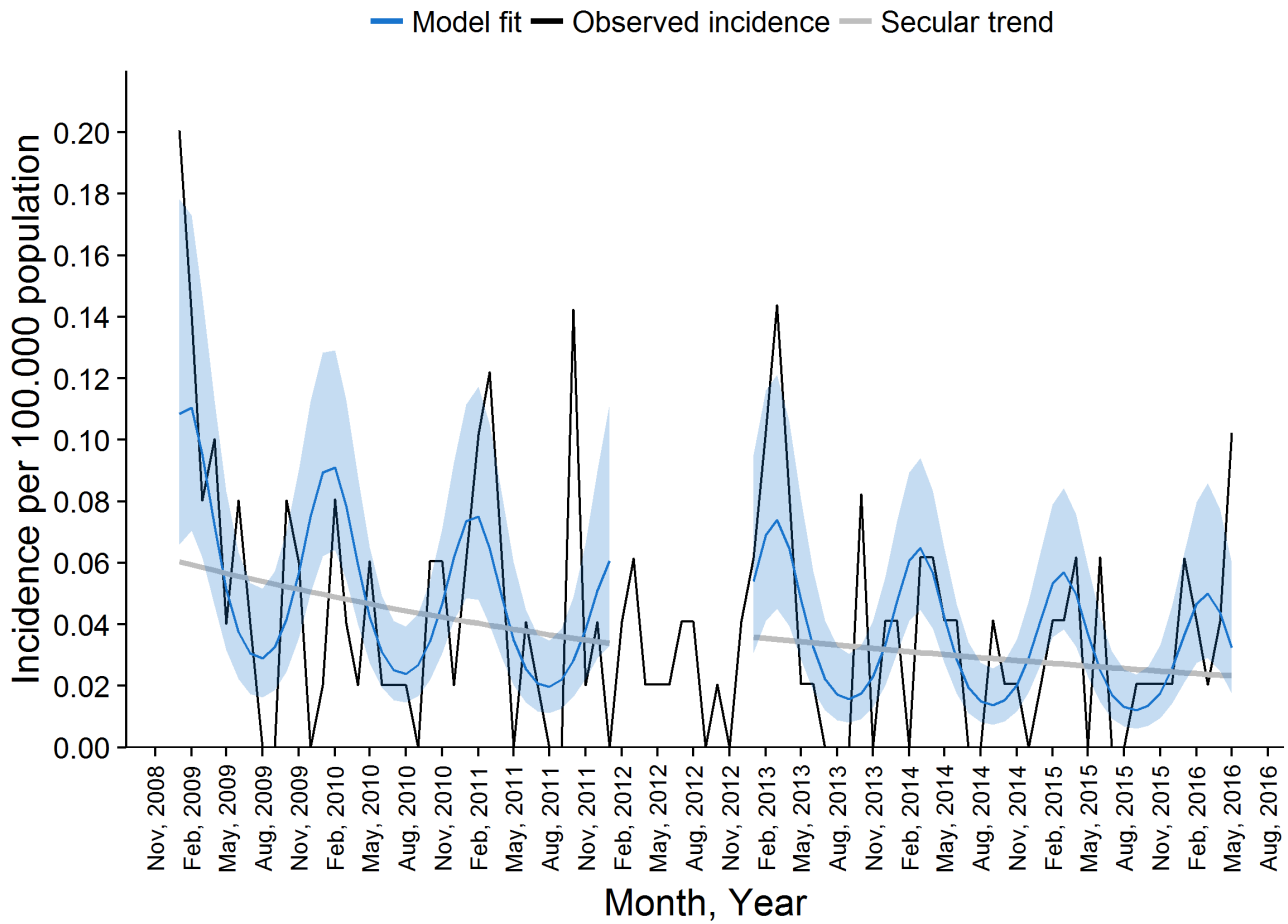
Results: Control outcome among ≥50 years old

Figure: Observed incidence/100,000 population, model fit with 95% CI and secular trend of the segmented regression model including 2012 as transition period. *Non PPV23 ST-IPD* among ≥50 years old, Austria, 2009 – 2016/05.



Results: Control outcome among 5–49 years old

Figure: Observed incidence/100,000 population, model fit with 95% CI and secular trend of the segmented regression model including 2012 as transition period. *PCV10 ST-IPD* among 5–49 years old, Austria, 2009 – 2016/05.



Summary

After the intervention:

- ≥50 years old
  - decline in PCV10 serotypes (intervention outcome)
  - no secular change in other serotype groups (control outcomes)
- 5–49 years old
  - no changes in any serotype group (control outcomes)

Conclusion

Indirect effect (herd immunity) among elders after intervention in <5 years old!

<5 years old

- Low number of cases, especially in subgroups (e.g. PCV10 ST-IPD)
  - ⇒ segmented regression not applicable
- „Molecular“ laboratory method (PCR) introduced in 2013
- Used for case finding and serotype determination
  - ⇒ possibly more cases found and more serotypes determined
  - ⇒ comparison of pre- and post-intervention period not applicable
  - ⇒ case counts are still low
- Use Zero-inflated and Hurdle models

Method: Zero-inflated vs Hurdle

- Developed to model count data with high occurence of zero counts
- Usable for Poisson and negative binomial distributed observations
- Zeros are interpreted as
  - structural zeros - due to some special structure of the data
  - sampling zeros - from sampling
- Zero-inflated: structural and sampling zeros
- Hurdle: structural zeros only

Model: Poisson version of zero-inflated and hurdle model

Zero-inflated model

$$P(Y_i = y_i) = \begin{cases} p_i + (1 - p_i)\, e^{-\mu_i}, & y_i = 0 \\ (1 - p_i)\, \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, & y_i > 0 \end{cases}$$

Hurdle model

$$P(Y_i = y_i) = \begin{cases} p_i, & y_i = 0 \\ (1 - p_i)\, \frac{e^{-\mu_i} \mu_i^{y_i}}{\left(1 - e^{-\mu_i}\right) y_i!}, & y_i > 0 \end{cases}$$

- p_i: probability of being excess zero - often modelled using logistic regression

Results using different models

Table: Model output (RRs and p-values) of different models using PCV10-ST IPD data including molecular typed cases on <5 years old, Austria, 2012 – 2016/05

variable	NB	p_{NB}	ZINB	p_{ZINB}	NBH	p_{NBH}	POIS	p_{POIS}	ZIP	p_{ZIP}	PH	p_{PH}
Intercept	0.26	0.002	0.27	<0.001	0.27	0.026	0.26	<0.001	0.27	<0.001	0.27	0.026
linear	0.95	0.035	0.97	0.082	0.97	0.255	0.95	0.018	0.97	0.082	0.97	0.255
$Intercept_{zero}$			0.00	0.620	1.26	0.732			0.00	0.537	1.26	0.732
$linear_{zero}$			3.05	0.672	0.95	0.091			2.70	0.577	0.95	0.091
AIC	81.38		79.44		88.04		79.92		77.45		86.04	

Conclusion, <5 years old

- Poisson models fitted data better than their negative binomial counterparts
  $\Rightarrow$ no significant overdispersion
- Ordinary poisson regression explained *Non PCV10 ST-IPD* and total IPD data best
  $\Rightarrow$ no, or less excess zeros compared to *PCV10-ST IPD*
- PCV10 serotypes: secular decline (intervention outcome)
- Non PCV10 serotypes: no changes (control outcome)

Thank you!
Any questions?

References

[1] A.C. Cameron and P.K. Trivedi. *Regression Analysis of Count Data*. Cambridge: Cambridge Univ. Press, 1998.

[2] D. Gillings, D. Makuc, and E. Siegel. "Analysis of interrupted time series mortality trends: an example to evaluate regionalized perinatal care". eng. In: *American Journal of Public Health* 71.1 (Jan. 1981), pp. 38–46. ISSN: 0090-0036.

[3] A. K. Wagner et al. "Segmented regression analysis of interrupted time series studies in medication use research". eng. In: *Journal of Clinical Pharmacy and Therapeutics* 27.4 (Aug. 2002), pp. 299–309. ISSN: 0269-4727.

[4] Robert B. Penfold and Fang Zhang. "Use of Interrupted Time Series Analysis in Evaluating Health Care Quality Improvements". In: *Academic Pediatrics* 13.6, Supplement (2013). Quality Improvement in Pediatric Health Care, S38 –S44. ISSN: 1876-2859. DOI: http://dx.doi.org/10.1016/j.acap.2013.08.002. URL: http://www.sciencedirect.com/science/article/pii/S1876285913002106.

[5] C. E. Rose et al. "On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data". eng. In: *Journal of Biopharmaceutical Statistics* 16.4 (2006), pp. 463–481. ISSN: 1054-3406. DOI: 10.1080/10543400600719384.

[6] Mei-Chen Hu, Martina Pavlicova, and Edward V. Nunes. "Zero-inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial". In: *The American journal of drug and alcohol abuse* 37.5 (Sept. 2011). ISSN: 0095-2990. DOI: 10.3109/00952990.2011.597280. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3238139/ (visited on 12/19/2016).