Download materials for today’s workshop from:
https://github.com/weinbergerlab/ISPPD-workshop

11th ISPPD Workshop #2

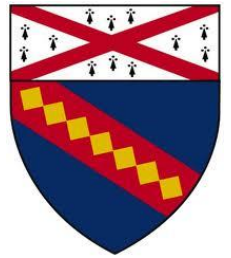
Evaluating Vaccine Impact using Time Series Data

Dan Weinberger, PhD
Kayoko Shioda, DVM, MPH

Epidemiology of Microbial Diseases
Yale School of Public Health

April 15, 2018

Yale



Conflicts: DMW had received consulting fees from Pfizer and Affinivax

Today

Lecture #1

- Brief intro to counterfactuals
- Methods to calculate counterfactuals (**Part 1**)
  - Pre-post comparison
  - Linear trend model
  - Interrupted time series

Lab #1

- Interrupted time series model

Lecture #2

- Methods to calculate counterfactuals (**Part 2**)
  - Control variables
  - Synthetic control

Lab #2

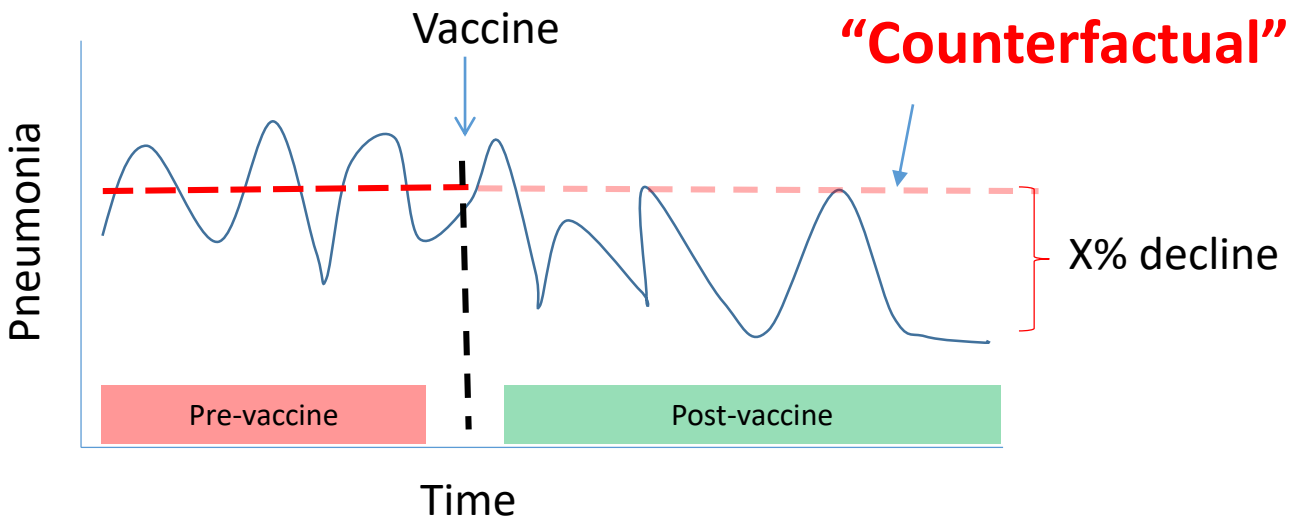
- Synthetic control

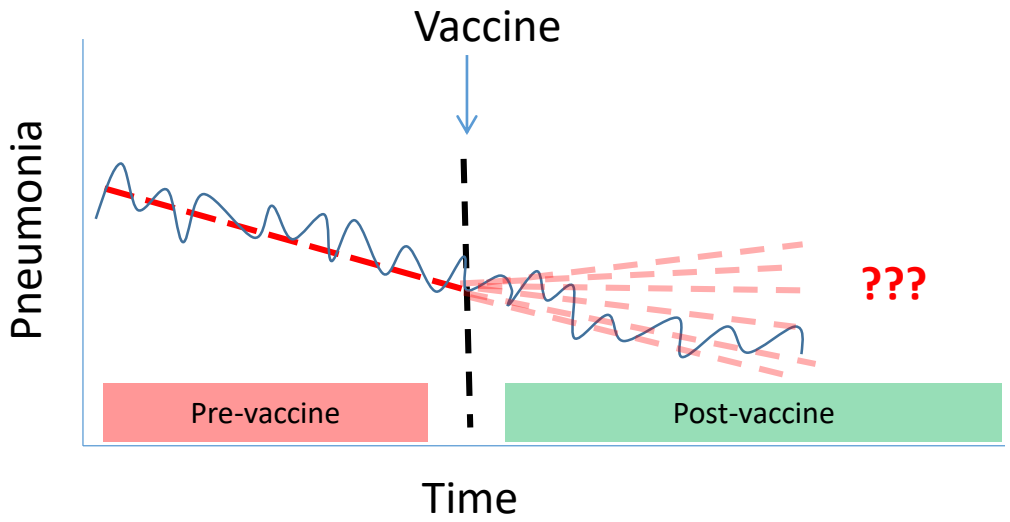
Discussion and Q&A

Lecture 1

Counterfactuals

Evaluating the impact of PCVs from time series





- **What we need to know**: What would have happened without vaccine (the counterfactual)

- Estimating this quantity is a major challenge and relies on various assumptions

Key questions

- Is my disease changing over time in absence of vaccine?
- How quickly is disease changing in absence of vaccine?
- Is PCV having an impact on rates?

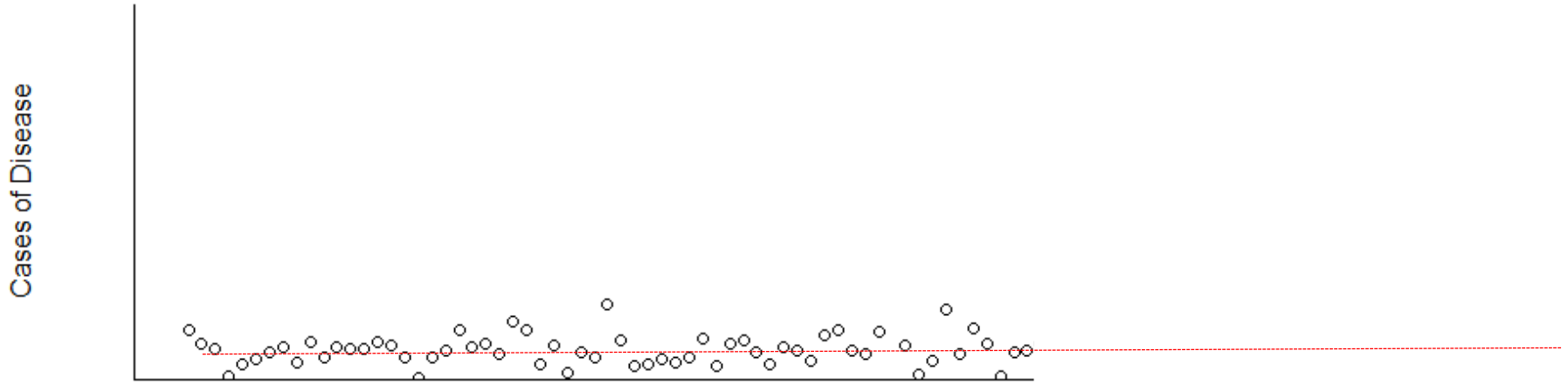
How do we establish a counterfactual?

- **Most common:** Some type of regression or time series model

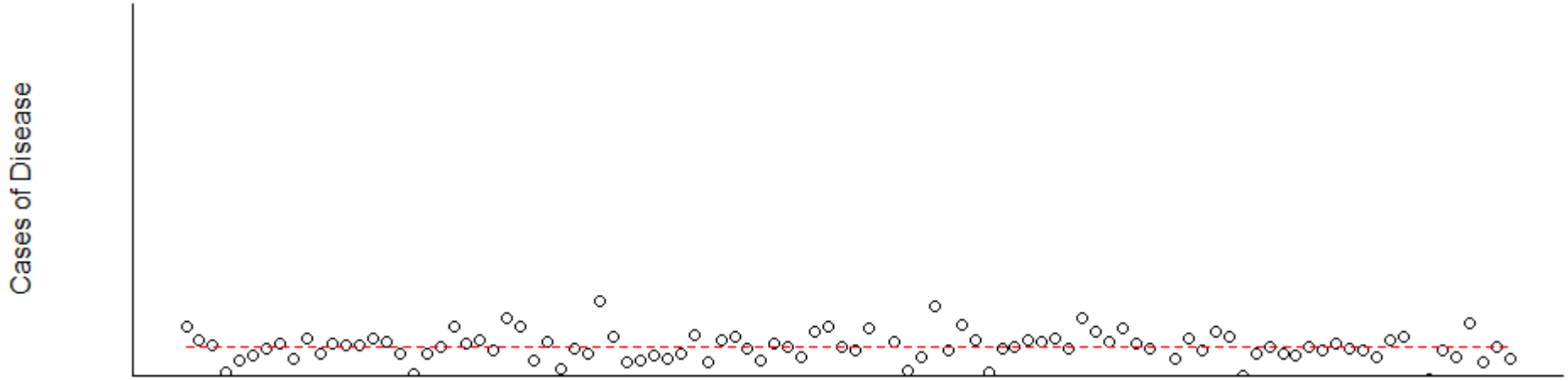
- **More complicated:** Simulation model/transmission model

Pre-post analysis

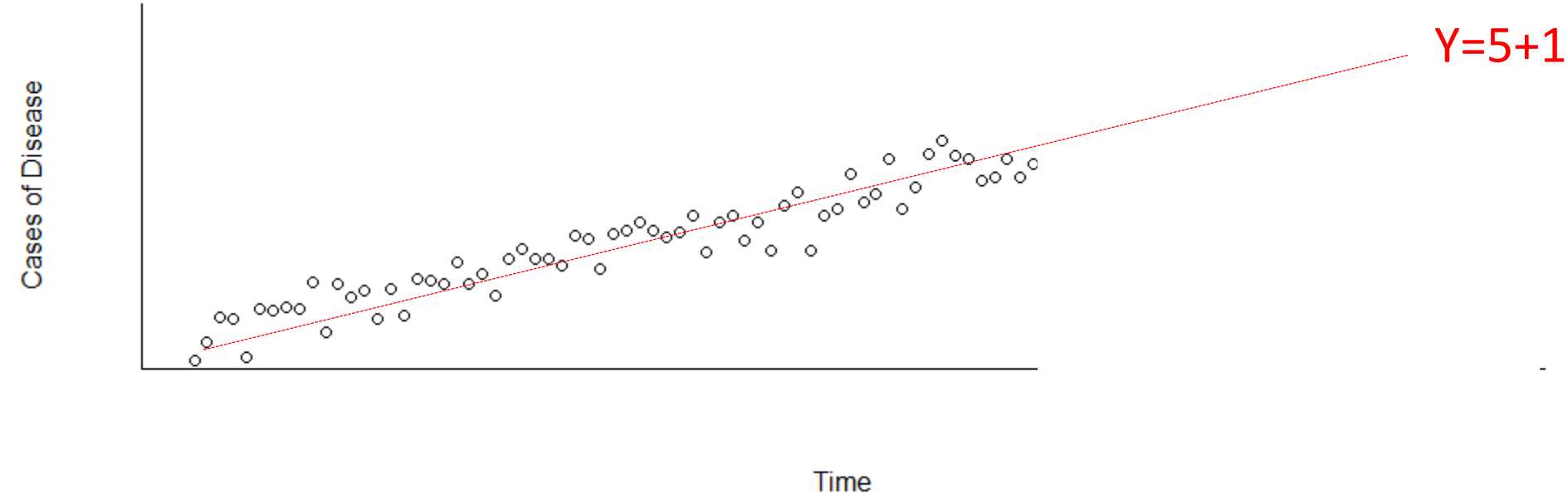
- Simplest case: stationary data ; no trends, no seasonality
- Test whether mean number of cases declines post-PCV

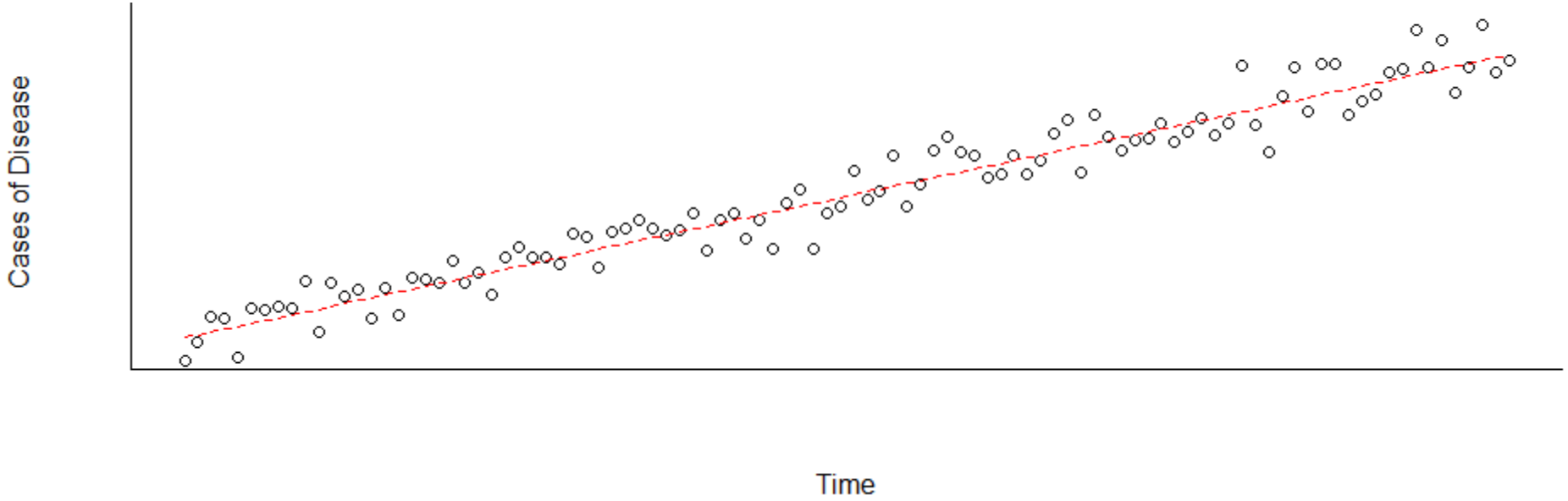


Simple model:
Y=5+e



SIMPLE TRENDS





More complicated patterns to incorporate

- Seasonality
- Non-linear patterns (e.g., polynomials)

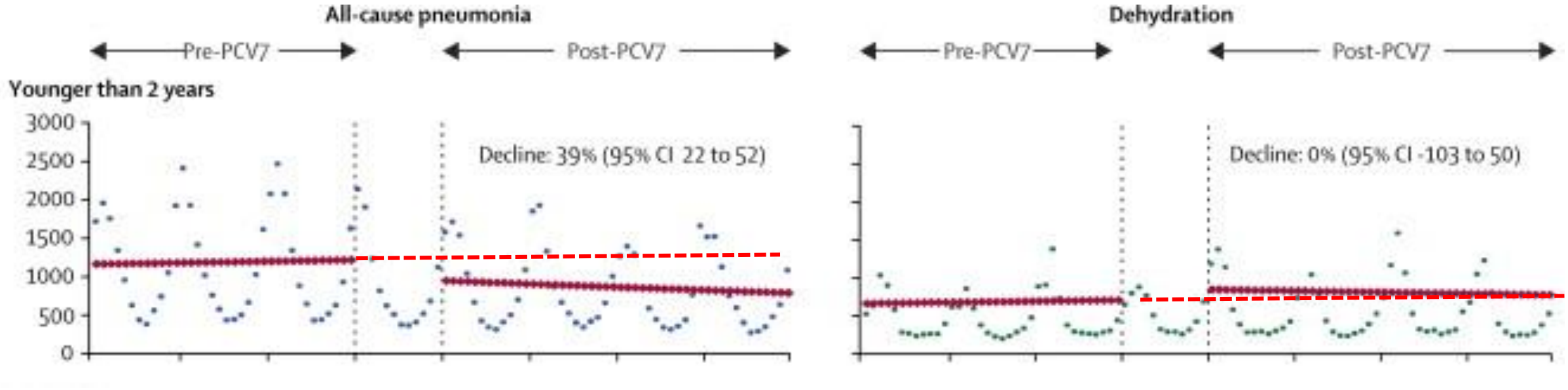
Evaluating changes in trend

- **Method 1:** Fit model to pre-vaccine period and extrapolate to post-vaccine period. Compare observed vs expected values (Rate ratio) at each time point
- **Method 2: Interrupted time series:** Fit trend model to entire time series and use interaction terms to test for change in trend

With both approaches: assume that trend occurring in the pre-vaccine period would have continued into the post-vaccine period

Variations on this theme: ARIMA models, Holt-Winter model

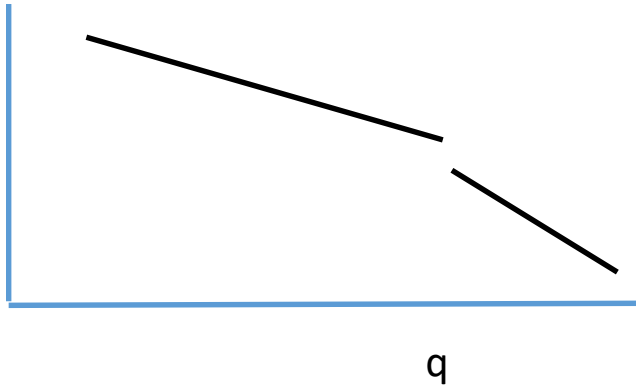
Example of interrupted time series:
trends in pneumonia following introduction of PCV7 in the US



Grijalva et al, The Lancet 2006

Pneumonia declines 39% *compared to what would have been expected if not vaccine was introduced*

Testing for a change in trend: Interrupted time series (ITS)



Does slope significantly change at time q

- Pneumonia rate = exp(b+at+c**z**+d**z**t)
- z is a dummy variable
  - 0 before time t, 1 after
  - Allows for a different slope before (a) and after (a+d) time t

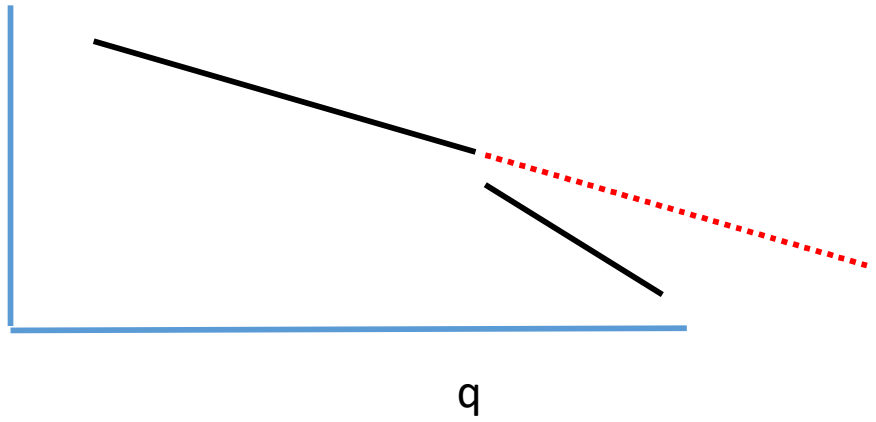
*Is coefficient for the interaction term “d” significant?

*Can evaluate importance of interaction with p value of interaction; Likelihood ratio test; AIC score.

Counterfactual from ITS

Does slope significantly change at time q

- Predicted value at time t: Yt=exp(b+ax+cz+dxz)
- Counterfactual at time t: Yt=exp(b+ax)



Difference or ratio between observed and counterfactual lines gives the “Vaccine impact” (rate ratio or rate difference)

What could go WRONG?

- Epidemic before or after vaccine introduction (biases slope estimates)
  - Ie 2009 pandemic, then introduce PCV in 2010
- Insufficient data in pre- or post-period to accurately estimate trend
- Unrelated changes that coincide with vaccine introduction
- Delayed rollout of vaccine/low uptake
- Many, many other issues that you can’t predict…

Sensitivity analyses you should always do

- Never trust your main analyses without “pressure testing” it
- Try different intervention dates—how does it influence your estimate?
  - Change point analysis can be thought of as a sensitivity analysis for ITS
- Leaving out different seasons when fitting, see if it changes the answer
  - Even better: bootstrap seasons to test robustness
- Perform simulations to see how likely it is that you would detect a decline
  - https://weinbergerlab.shinyapps.io/shinyplay03/

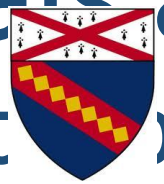
Steps for evaluating change in trend

1. Define “baseline”, transition periods
2. Determine whether there are any trends or patterns in the baseline period
   - Seasonality, etc
3. Fit a model to your baseline data using regression
   - Be wary of over-fitting (use AIC)
4. Compare test and reference periods

THINK ABOUT WHAT COULD GO WRONG:

- Identify controls!
- Do sensitivity analyses!
  - i.e., leave out one season at a time; try different intervention times

Lab 1

Materials at:
https://git  .com/weinbergerlab/ISPPD-workshop

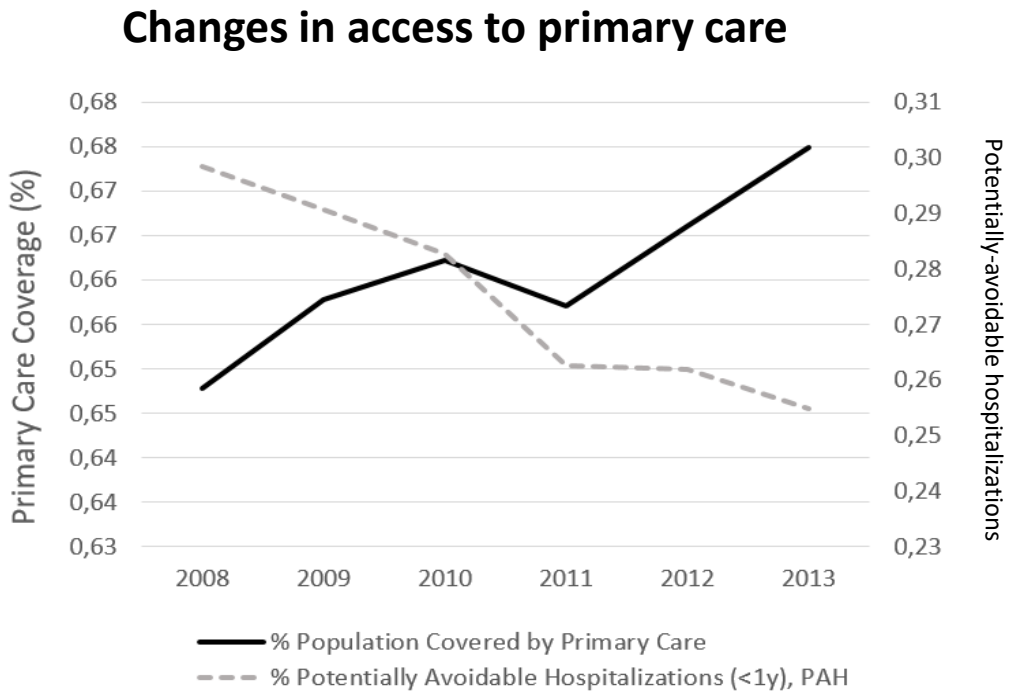
Lab 1

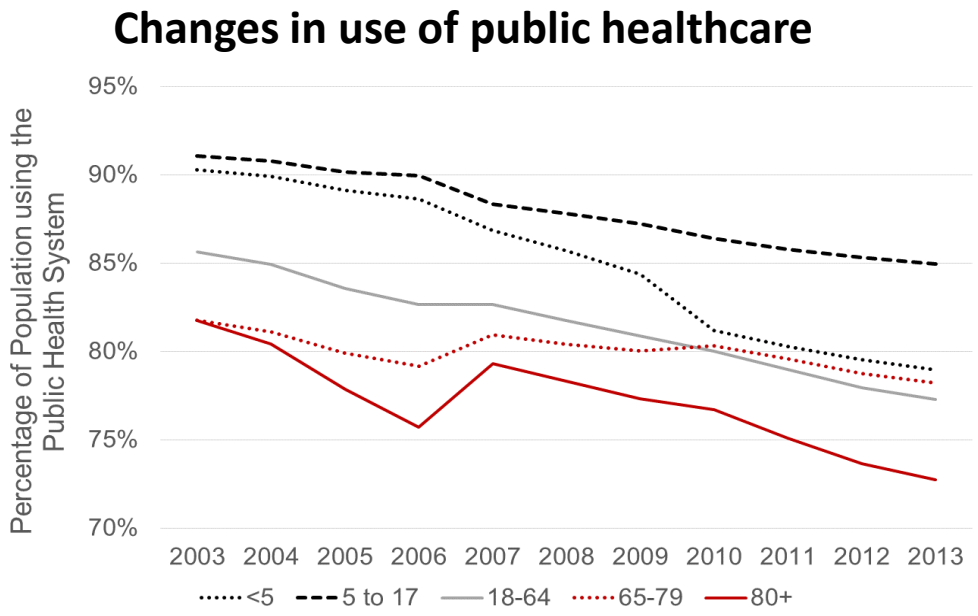
1. Fit models to pre-vaccine data and extrapolate trends
2. Testing for a change in trend using a simple interrupted time series: dummy variable for time period, trend, and an interaction between trend and time period
3. Estimate the counterfactuals with different methods

Lecture 2

Synthetic control

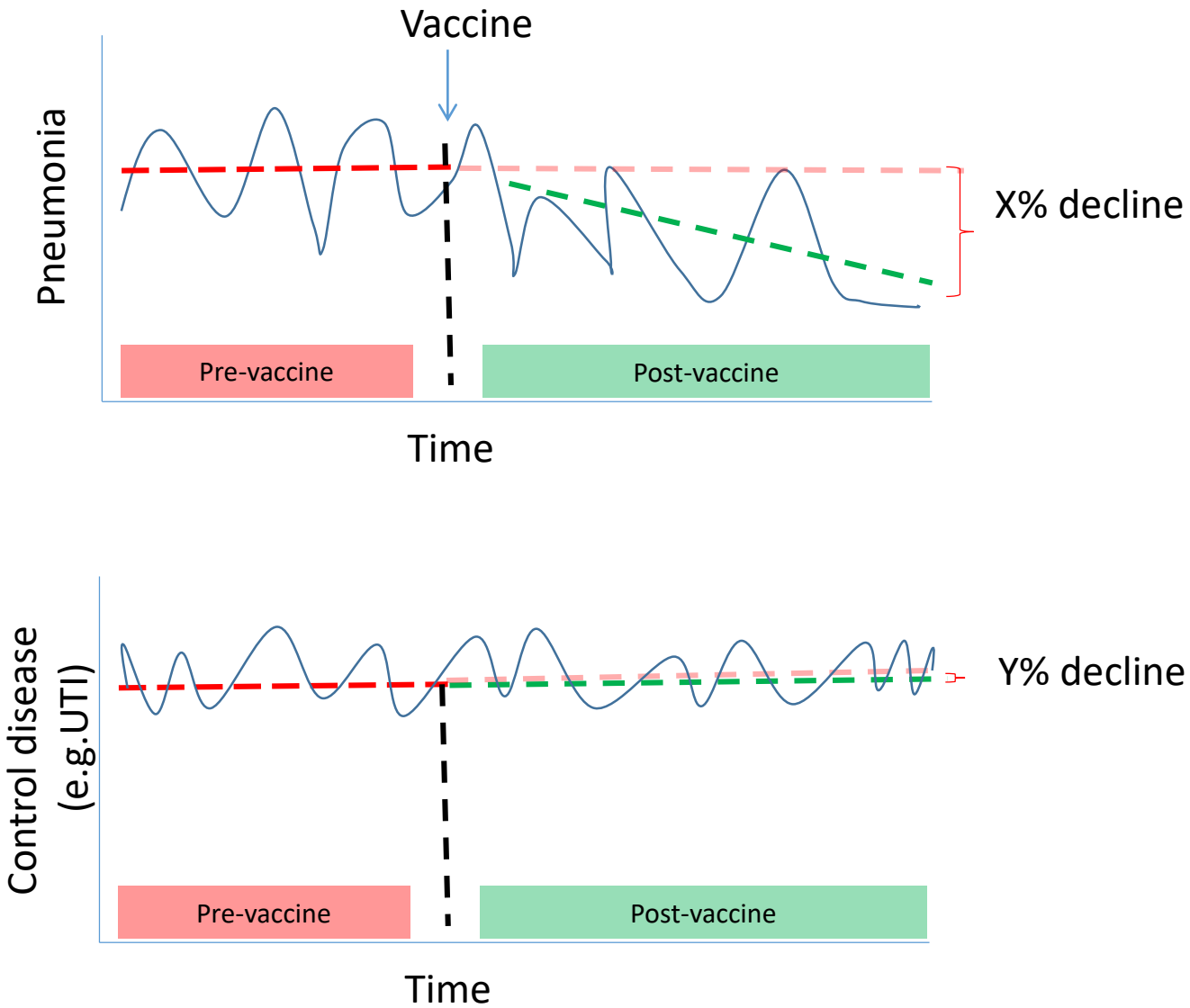
Many factors aside from vaccination can influence disease rates





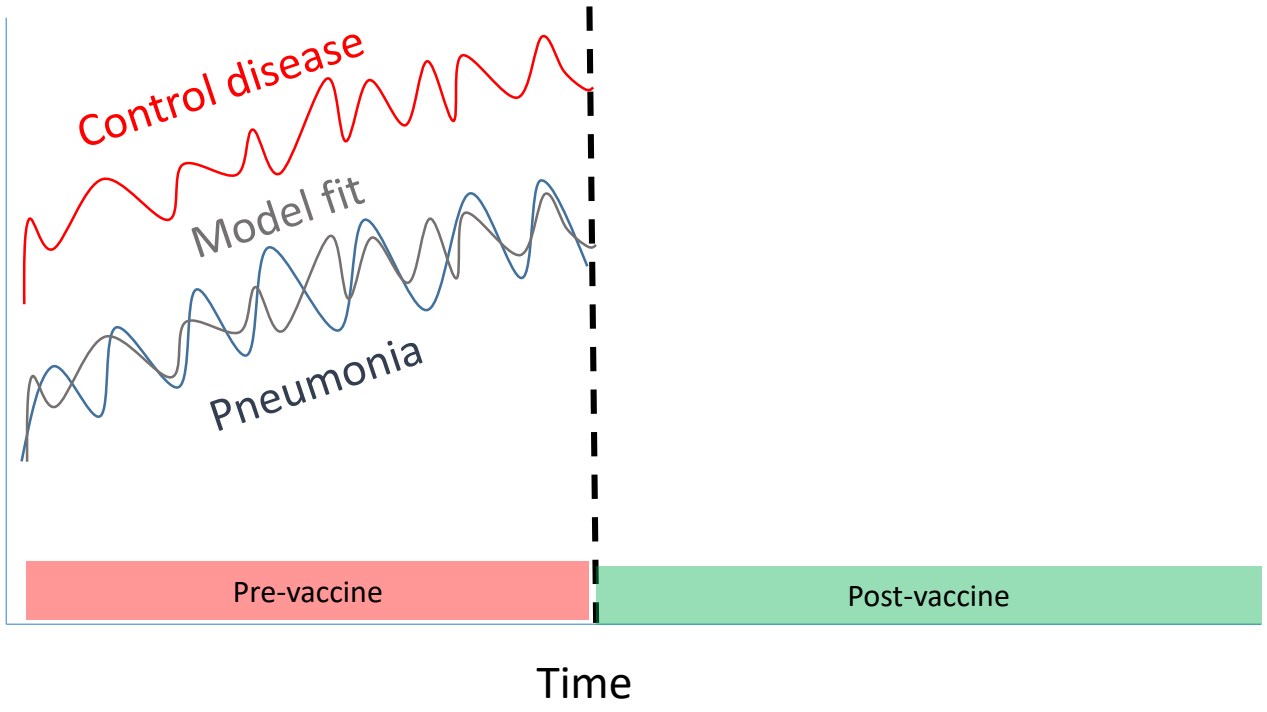
Cynthia Schuck-Paim, *Vaccine 2017*

Use control diseases to detect/adjust for unrelated trends



- Often used qualitatively
  - "Pneumonia declines but UTI is stable"
- Can be used quantitatively
  - "Effect of PCV against pneumonia is X%-Y%"

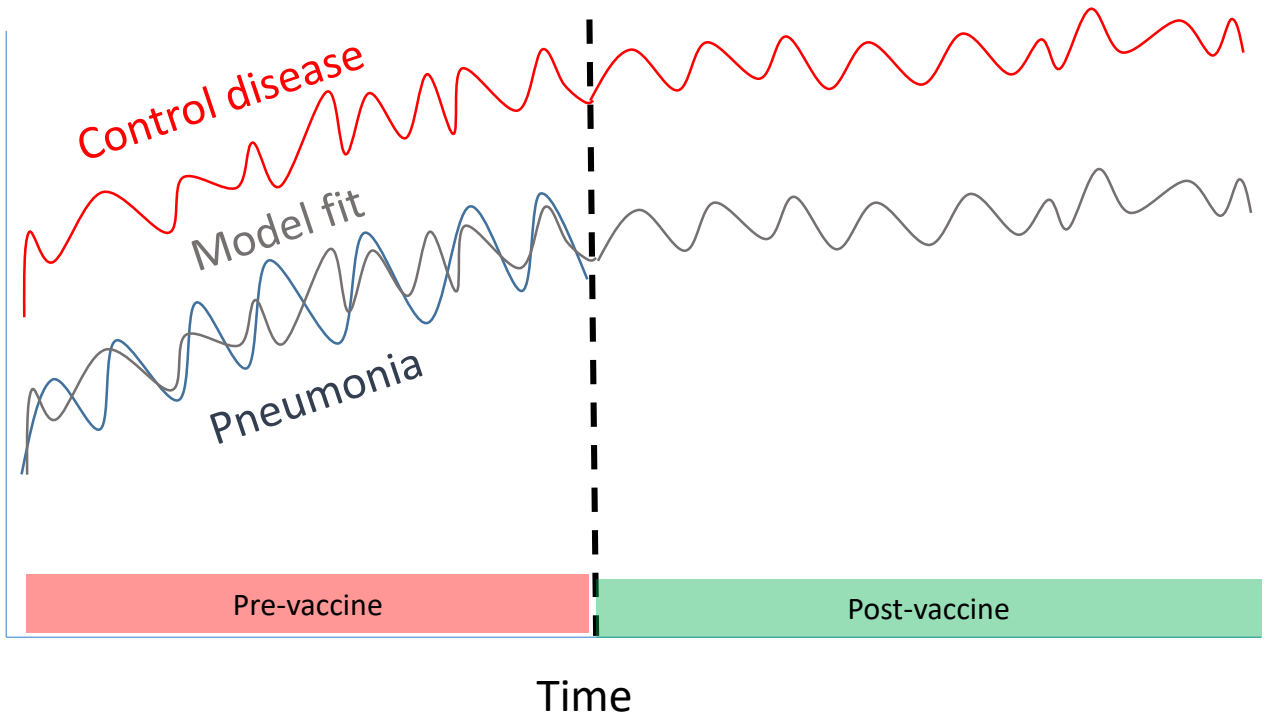
How does it work?



Step 1: Fit a regression model using data from the pre-vaccine data to establish a relationship between pneumonia and a control disease

E.g., log(pneumonia)= b0 + b1*log(control disease)

How does it work?



Step 2: Plug in observed values for control disease from post-vaccine period to get an estimate for what counts of pneumonia would be

Key Assumptions

- Relationship between pneumonia and control is stable over time and only change is due to the vaccine
  - Violated if there is an intervention that influences the control
- Assumes control disease shares important non-vaccine trends with pneumonia

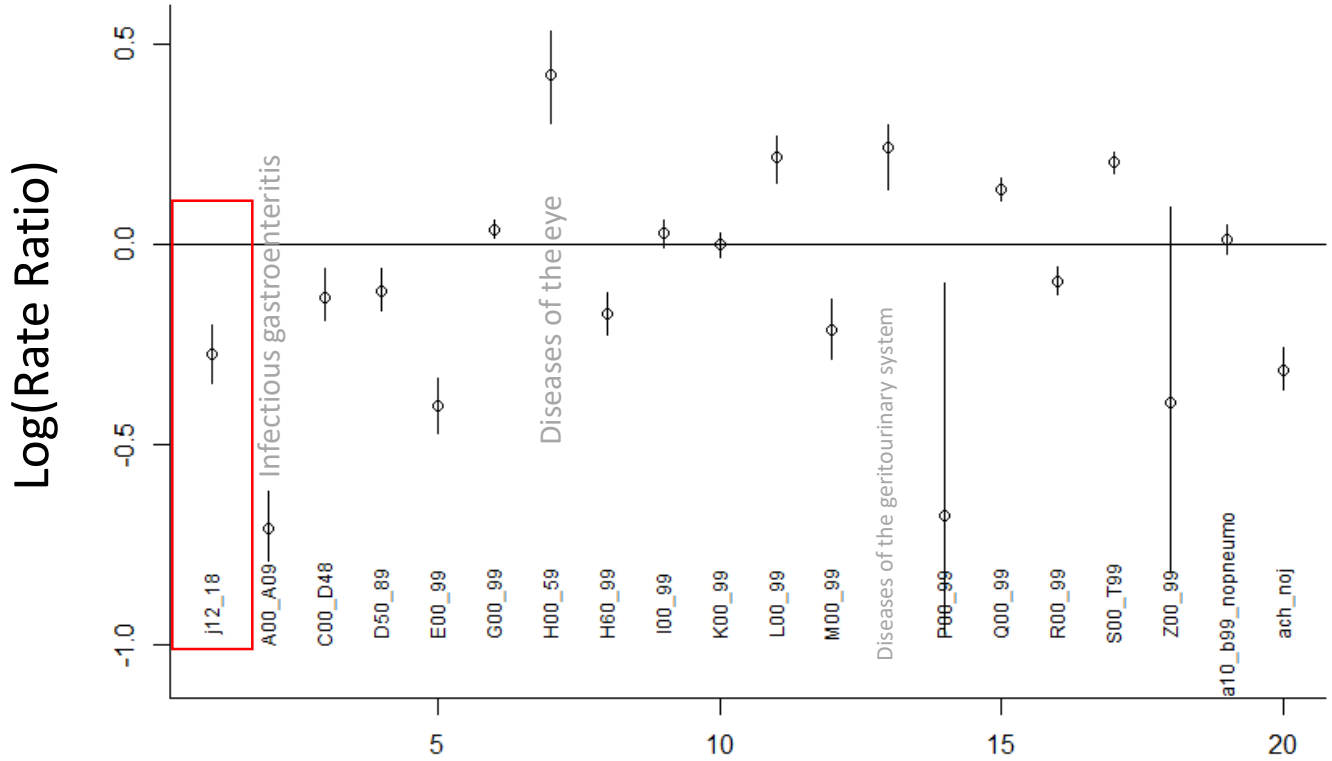
What is a good control for pneumonia?

What has been used as a control for PCV impact against pneumonia?

- Urinary tract infections
  - Acute event
  - Definitely not influenced by vaccine
  - Only influences some age groups
  - Different etiology
- Fractures
  - Might capture some broad healthcare utilization patterns (?)
  - Definitely not influenced by vaccine
  - Very different risk factors, causal mechanisms from pneumonia
- Bronchiolitis
  - Closest in etiology to pneumonia
  - Possibly influenced by the vaccine
  - Only occurs in certain age groups

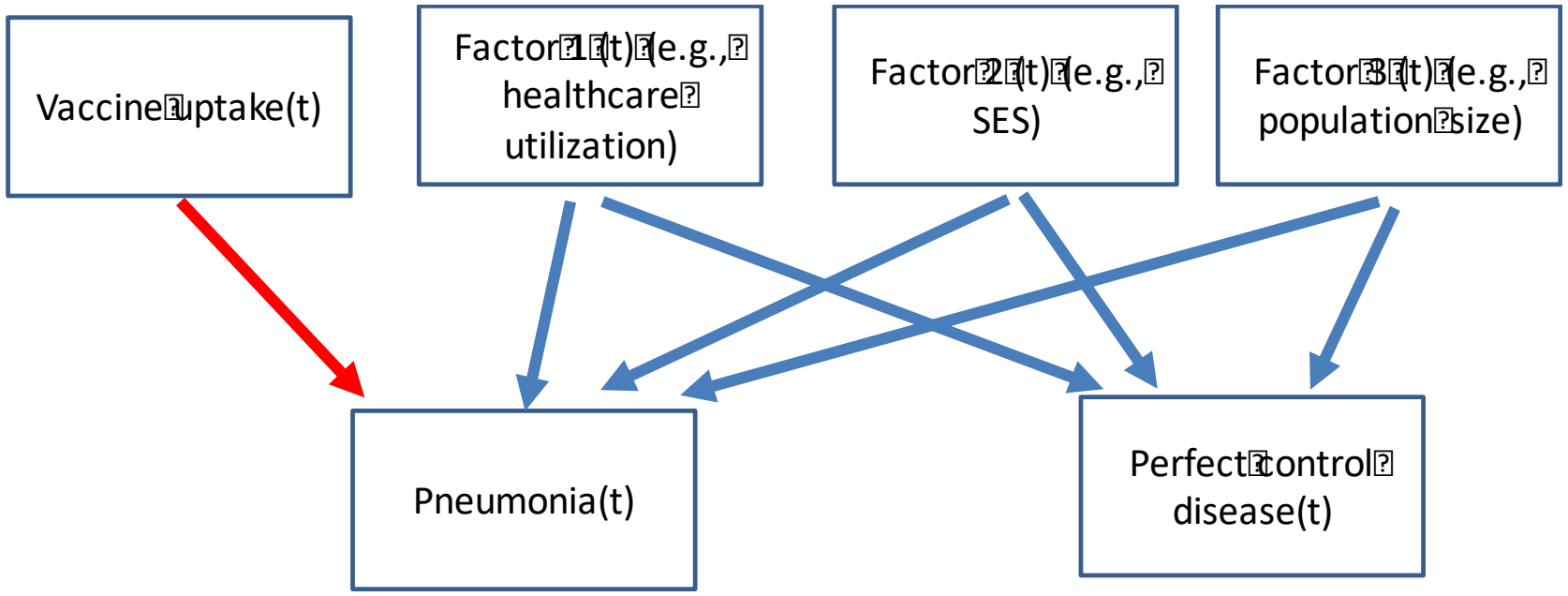
The challenge: Which control should we choose?

Changes in different disease categories post-PCV10



**Choosing a single comparator/control is risky—composites are more robust

The ideal control: Shares all causal factors, but is not influenced by vaccine



Regression: E(pneumonia cases_t)= b0 + b1*Perfect_control_t

The problem: how to identify a good control

Principles for selecting candidate controls

- Exclude any that could plausibly be influenced by the vaccine (e.g. pneumococcal/streptococcal septicemia)
- Relationship should be stable over time (e.g. exclude diarrhea following rotavax)
- Exclude covariates with sparse data (<10 cases/month on average)

Letting the data select controls

- Method developed by Google for website analytics (Brodersen)
- Select large number of candidate controls *a priori*
- Fit regression model to pre-vaccine time series
  - Weight the candidate controls using Bayesian variable selection
- Generate counterfactual for post-vaccine period from model

Christian Bruhn

PNAS

Estimating the population-level impact of vaccines using synthetic controls

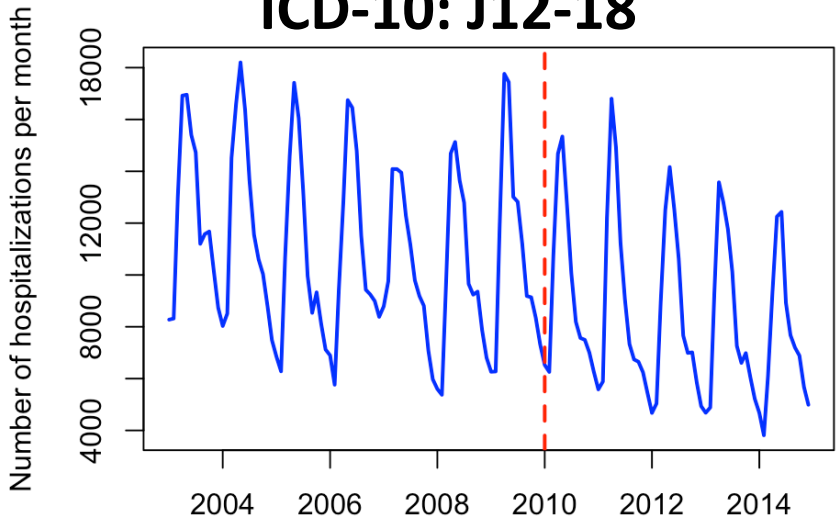
Christian A. W. Bruhn[a], Stephen Hetterich[b], Cynthia Schuck-Paim[b], Esra Kürüm[a,c], Robert J. Taylor[b], Roger Lustig[b], Eugene D. Shapiro[a,d], Joshua L. Warren[a,e], Lone Simonsen[b,f,g], and Daniel M. Weinberger[a,1]

What does synthetic controls do?

Outcome

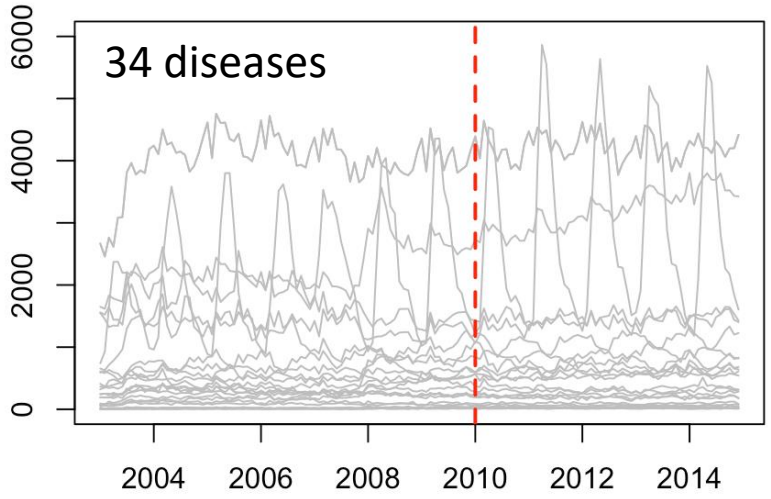
(all-cause pneumonia hospitalizations)

ICD-10: J12-18



Control diseases

Various ICD-10 codes



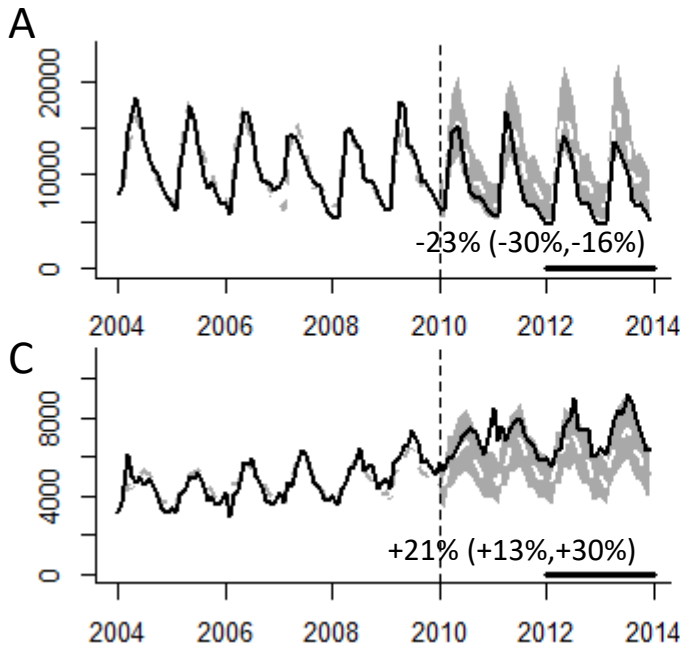
(Children <12 mo in Brazil is used as an example)

What does synthetic controls do?

- **Fit** regression model to **pre-vaccine** data only
  - Test different control variables alone or in combination
  - In forward or backward variable selection, you would drop less important variables
  - With this approach (Bayesian variable selection), you never drop any variables, you just give them more weight
- Gives you a regression model with a set of controls that do the best job at explaining trends in pneumonia pre-PCV
- **Extrapolate** to post-PCV period based on changes in the control variables

Example: Pneumonia in Brazil

Adjust only for non-respiratory hospitalizations



<12 months

80+ years

-Synthetic controls do not affect estimates for <12month old children (no hidden biases detected)
-In adults >80, without synthetic control, would estimates a 21% increase, with synthetic control, no change

Example of control diseases

Grouping scheme	ICD-10	Description	Exclusions
ICD-10 chapters			
	C00-D48	Neoplasms	A40.3, B95
	D50-89	Diseases of blood and blood-forming organs and certain disorders involving the immune mechanism	
	E00-99	Endocrine, nutritional, metabolic disorders	
	G00-99_SY	Diseases of the nervous system	G00-G04
	H00-99_SY	Diseases of the ear and mastoid process	H10, H65, H66
	I00-99	Diseases of the circulatory system	
	K00-99	Diseases of the digestive system	
	L00-99	Diseases of the skin	
	M00-99	Diseases of the musculoskeletal system	
	N00-99	Disease	
	P00-99	Perinat	
	Q00-99		
		a	
	R00-99	Sympt	
		not	
	S00-T99	Injury,	
	U00-99	Codes	
	V00-Y99	Externa	
	Z00-99	Factors	

Key assumptions:

1. Control diseases are **NOT affected** by the vaccine
2. Relationships between pneumonia and other diseases **would not change** over time, if we did **not** introduce PCVs

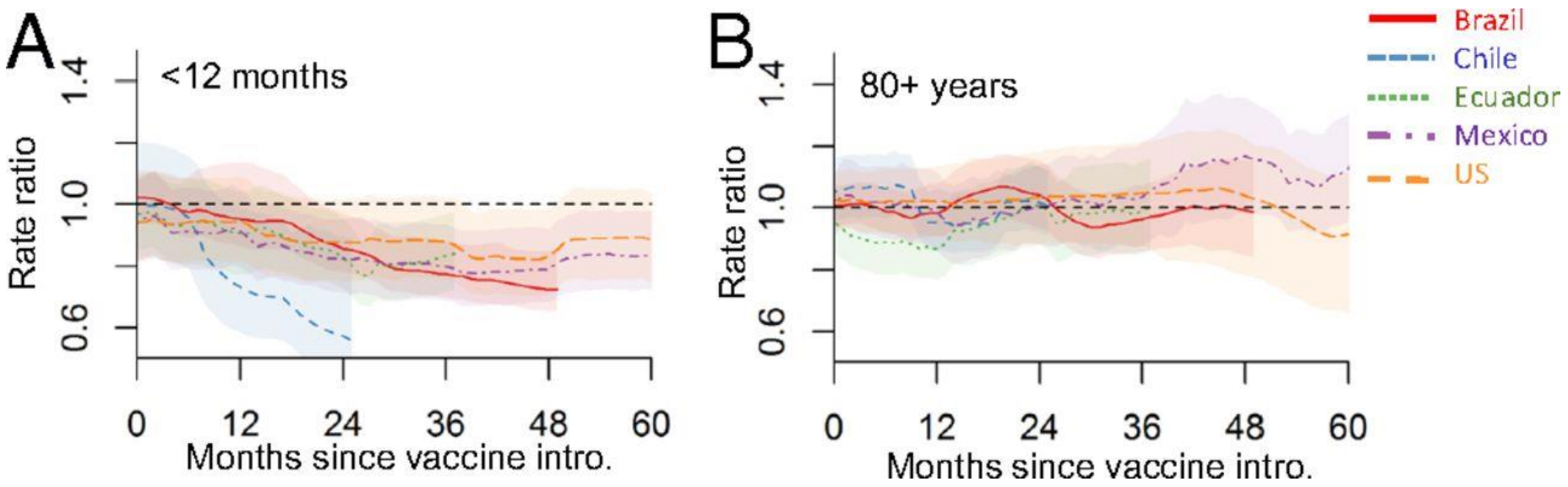
Which disease categories receive the most weight as controls?

- Some consistency in which controls receive most weight
- Method allows for flexibility between age groups and locations

80+y

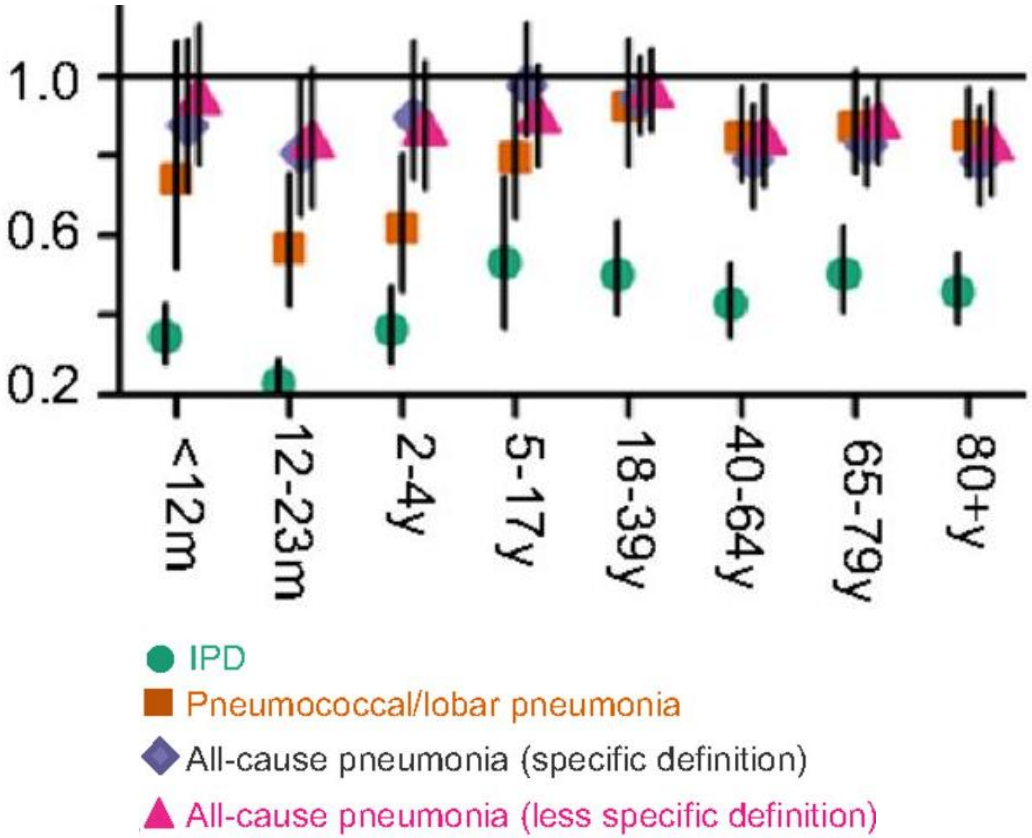
country.id	Brazil	Chile	Ecuador	Mexico
A10_B99_nopneumo	0.0729	0.1057	0.0117	0.0626
A41	0.7246	0.1386	0.0234	0.0258
ach_noj	0.1194	0.4934	0.9649	0.1014
C00_D48	0.07	0.9387	0.2425	0.0315
cJ20_J22	0.0175	0.015	0.6999	0.7706
D50_89	0.0488	0.2501	0.0158	0.0207
E00_99	0.079	0.0407	0.038	0.5002
E10_14	0.117	0.0358	0.0348	0.4404
E40_46	0.036	NA	NA	NA
G00_99_SY	0.021	0.0188	0.0178	0.023
H00_99_SY	0.1805	0.0219	0.026	0.0328
I00_99	0.6292	0.608	0.051	0.0452
I60_64	0.1552	0.0323	0.0615	0.0248
K00_99	0.0535	0.0345	0.0621	0.0848
K35	0.0153	0.0122	0.03	NA
K80	0.1365	0.0301	0.0212	0.0245
L00_99	0.1427	0.0347	0.0185	0.0411
M00_99	0.0306	0.0689	0.0359	0.0252
N00_99	0.0622	0.0474	0.0743	0.0334
N39	0.0869	0.0316	0.4343	0.0232
P00_99	0.015	NA	NA	NA
pandemic	0.0106	0.0304	0.0128	NA
Q00_99	0.032	NA	NA	NA
S00_T99	0.1006	0.034	0.0344	0.0562
Z00_99	0.0283	0.0116	0.031	0.0397

Trajectory of declines in five countries

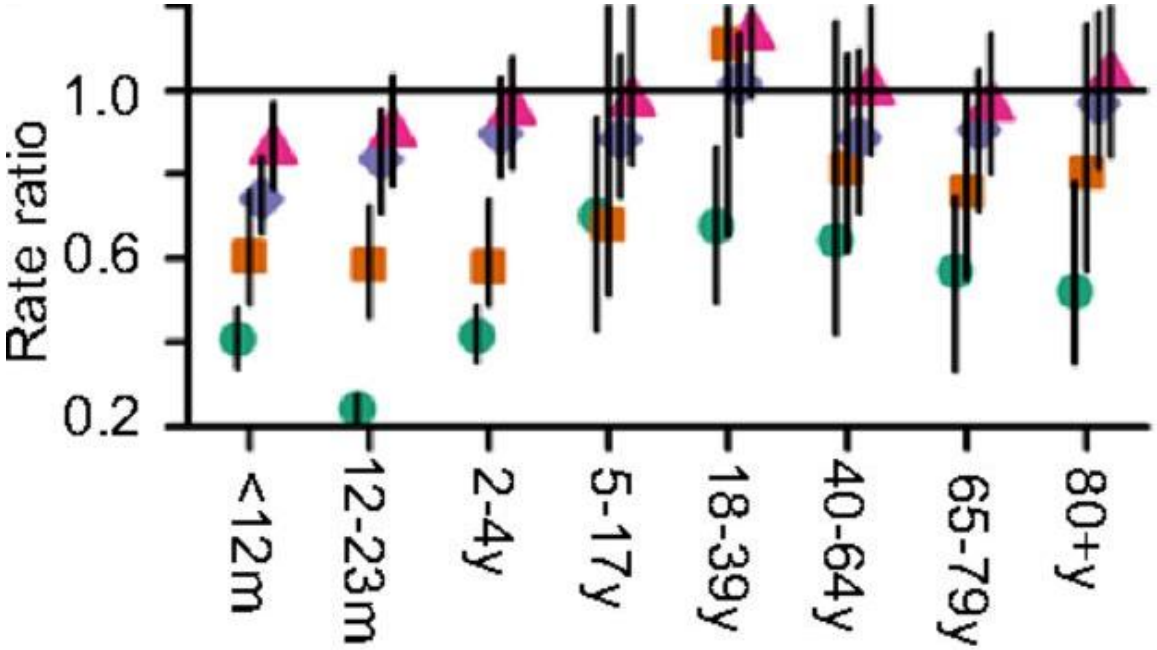


Impact of PCVs against outcomes of varying specificity

Simple trend adjustment



Synthetic control



Sensitivity analyses that are good to run

- If have 6 years of pre-vaccine data, fit model to first 5 years, estimate “rate ratio” for 5th year
  - Should be ~1
- Try dropping top 1,2,3 control variables; see if estimates change

Modifications to synthetic controls to simplify interpretation

- Fit model with each control disease individually
- Evaluate fit of model to pre-vaccine data to weight some more than others
- Average together estimates from individual models to get a consensus
- Make interactive visualizations

Demonstration of simpler approach

80+ year olds, Brazil pneumonia hospitalizations

Synthetic Controls: Pros and Cons

- Provides flexible and robust approach to estimate vaccine impact
- 2 strong assumptions
  - None of the controls are influenced by the vaccine
  - The relationship between pneumonia and the controls does not change over time
- Modifications needed for optimal use in small populations
- Doesn’t guarantee you will detect/adjust for all confounding, but it increases the chances of success

Extensions we are currently working on

- Modifications to use SC method with sparse data (see Kayoko Shioda's poster at ISPPD)
- More transparent way to measure importance of different control variables
- Method to pool results between different studies and increase credibility (See Alyssa Sbarra's talk at ISPPD)

Resources for using synthetic controls with administrative data

- Data and R scripts
  - https://github.com/weinbergerlab/synthetic-control

- Tutorial from Google
  - https://google.github.io/CausalImpact/CausalImpact.html

- Original Google Paper
  - https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/41854.pdf

Lab 2

Lab 2

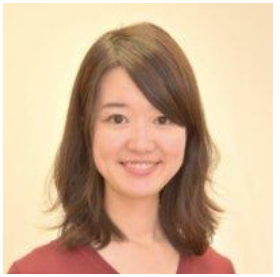
1. Review format of the data to use in the program
2. Estimate the impact of PCV in Brazil using the synthetic control analysis
3. Discuss the output and how to interpret it

Lecture 3

Alternative approaches

Synthetic controls with subnational data

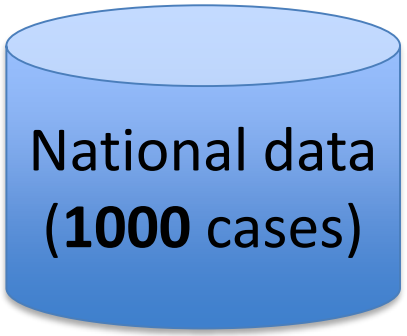
- With disaggregated data, more “noise” in the covariates
  - Might not be able to effectively adjust for shared trends
- Evaluate state-level variations in Brazil
- “Downsampling” simulation to test effect of population size



Kayoko Shioda

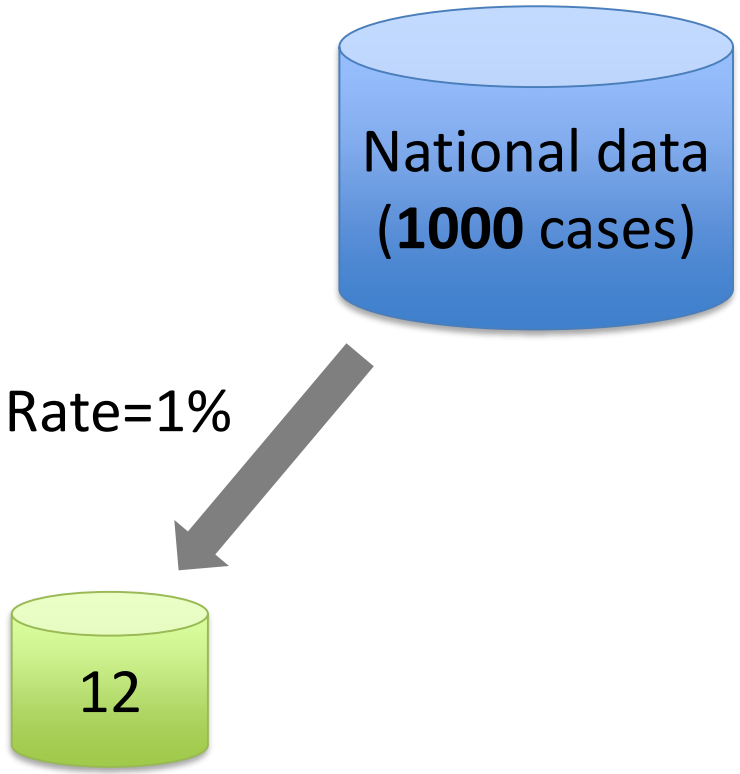
Down sampling analysis

Objective: To see how estimated RRs change by sample size



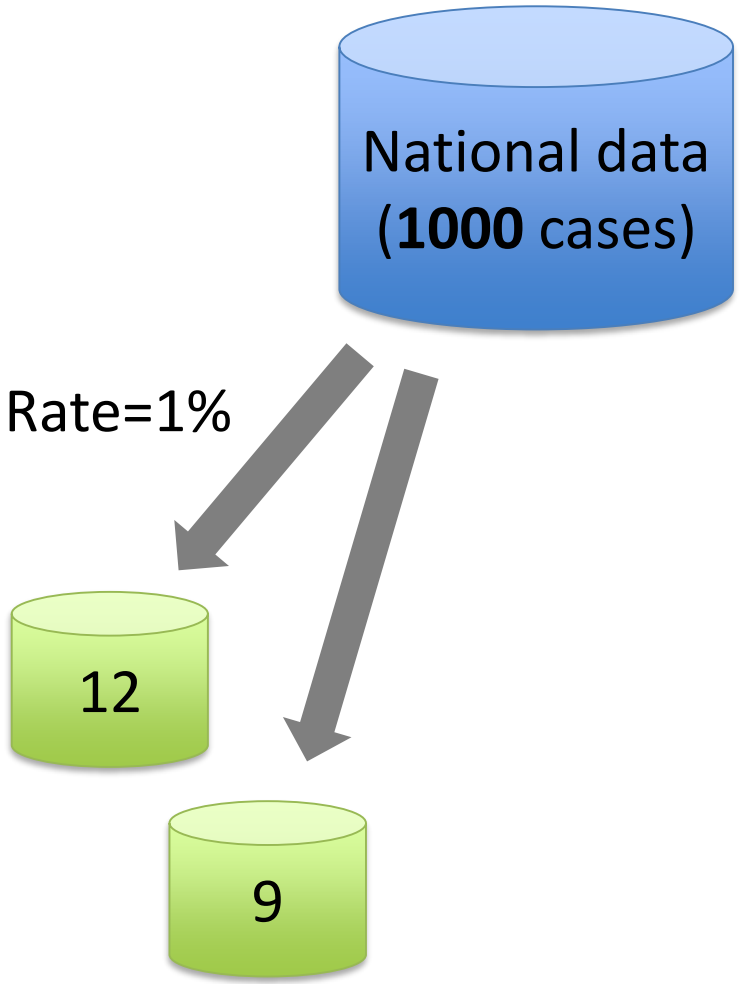
Down sampling analysis

Objective: To see how estimated RRs change by sample size



Down sampling analysis

Objective: To see how estimated RRs change by sample size



Down sampling analysis

Objective: To see how estimated RRs change by sample size

Down sampling analysis

Objective: To see how estimated RRs change by sample size

National data

Rates: 10%, 1%, 0.25%

Created 100 down-sampled datasets

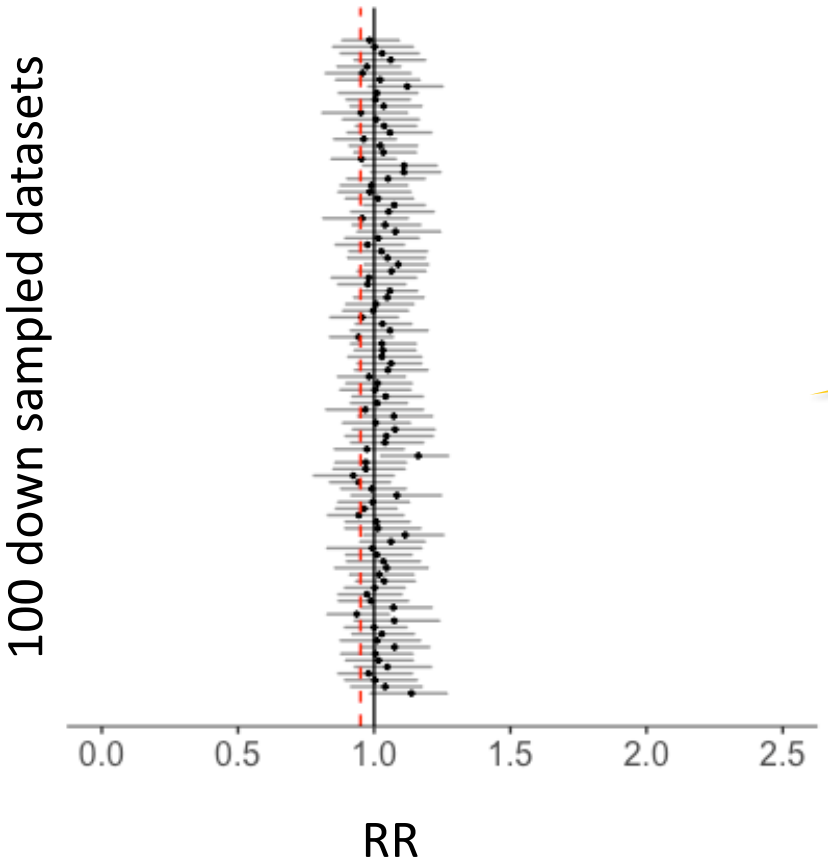
Run SC analysis for each down-sampled dataset
Calculate RRs for each dataset

E.g., RR = 9700 / 10000 = **0.97**
RR = 97 / 100 = **0.97**

RRs from 100 down-sampled datasets, 80+ yo

National estimate of RR = **0.95**
(represented by red dashed lines below)

Down sampling rate = 10%

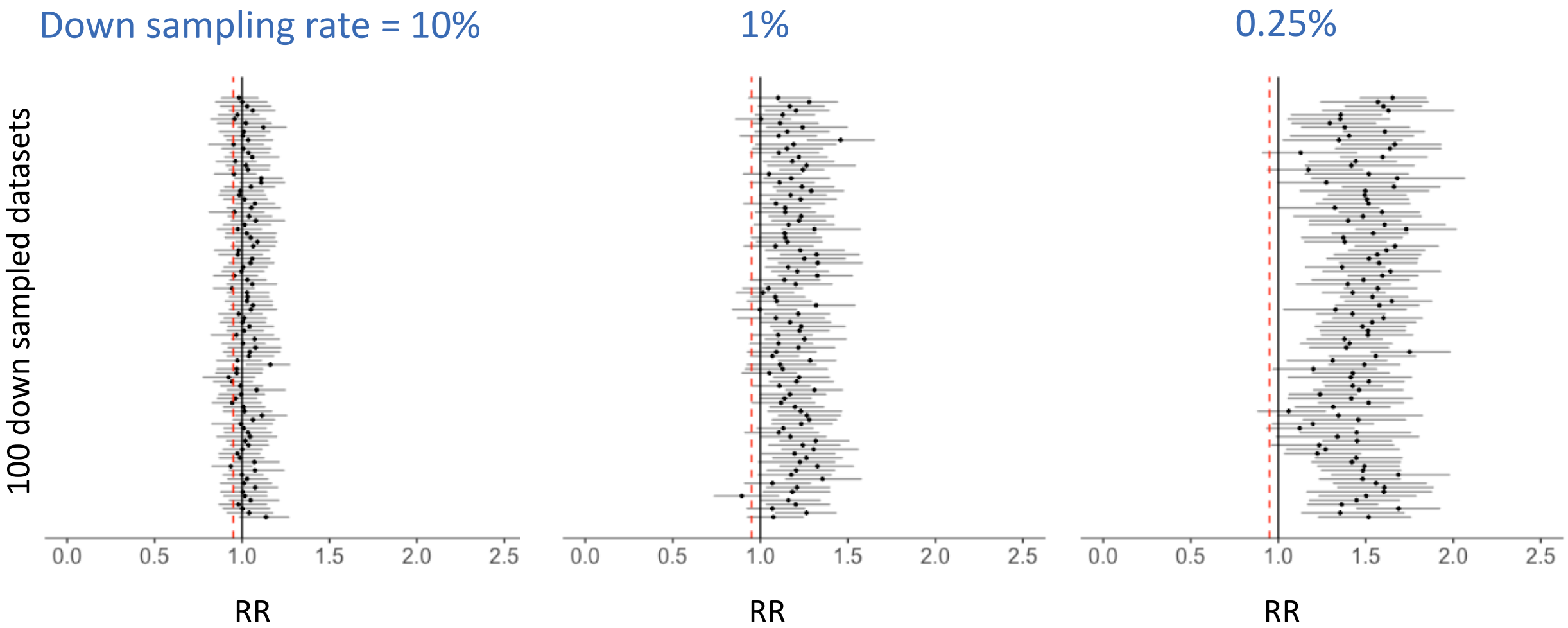


- **Red dashed line:** RR = 0.95 (national estimate)
- **Black line:** RR = 1

RRs from 100 down-sampled datasets, 80+ yo

National estimate of RR = **0.95**
(represented by red dashed lines below)

Not only CIs became wider but also RRs were biased away from the null



SC Model with Sparse Data on Control Diseases

- **Result:** SC model fails to generate reliable counterfactual when data on control diseases are sparse

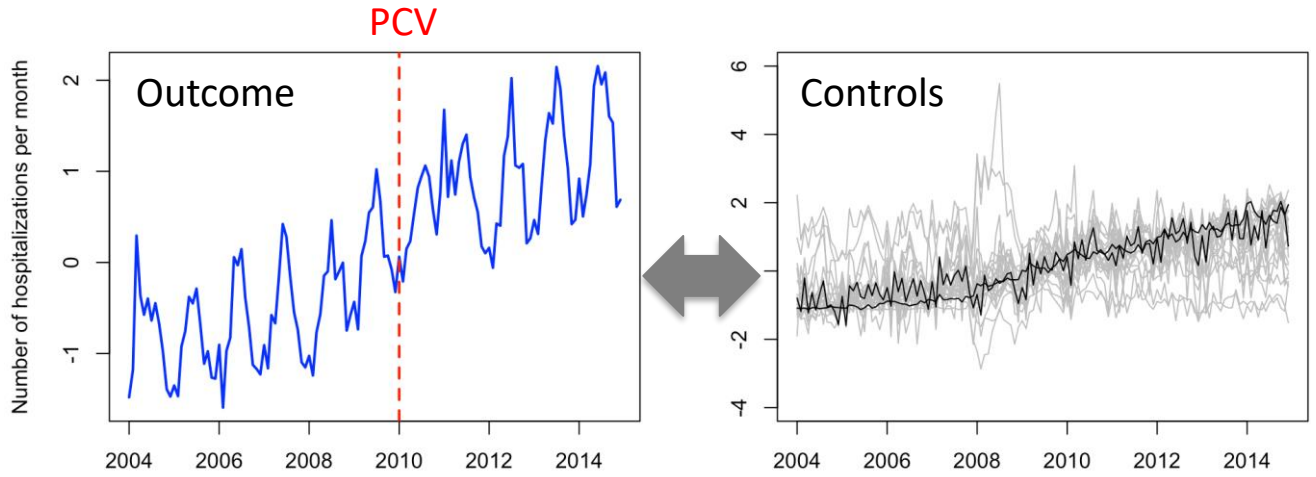
- **Why?**
  - Hard to assess correlations between the outcome and control diseases when data are noisy
  - As a result, SC model **fails to choose the best combination of control diseases or any control diseases**

SC Model with Sparse Data on Control Diseases

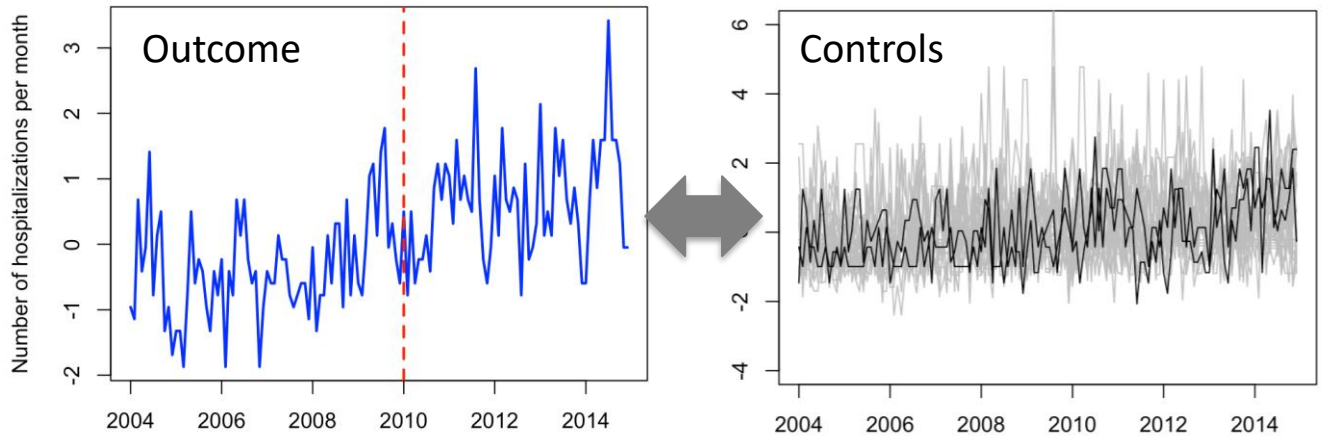
- **Result:** SC model fails to generate reliable counterfactual when data on control diseases are sparse

- **Why?**
  - Hard to assess correlations between the outcome and control diseases when data are noisy
  - As a result, SC model **fails to choose the best combination of control diseases or any control diseases**

National:

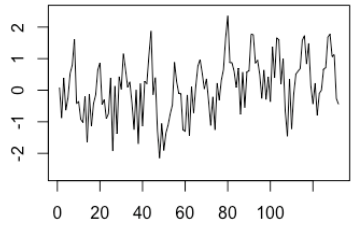


Down-sampled (0.25%):

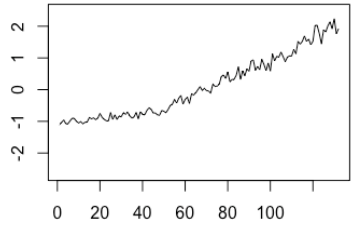


New Approach

Control disease (I00_99)

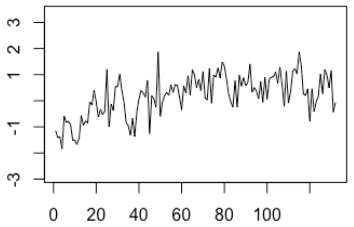


Control disease (A41)



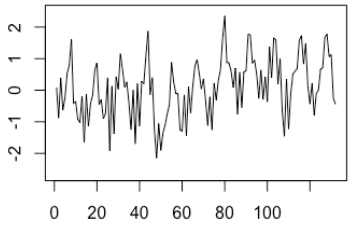
...

Control disease (G00_99_SY)

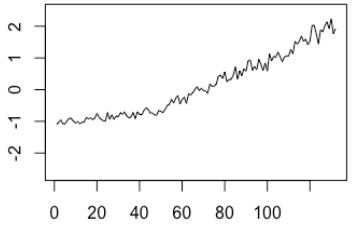


New Approach

Control disease (I00_99)

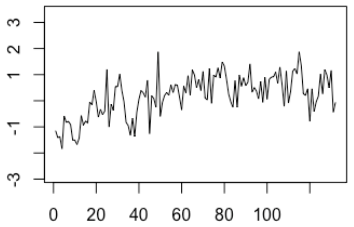


Control disease (A41)



...

Control disease (G00_99_SY)



Key assumptions:

1. Control diseases are **NOT affected** by PCVs
2. Relationships between pneumonia and control diseases **would not change** over time, if we did **not** introduce PCVs

(Same as the original SC model)

New Approach

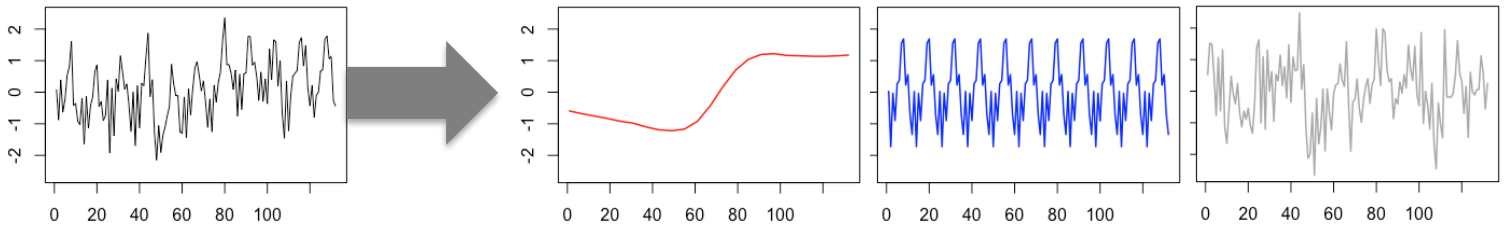
Step 1: Decomposition

Control disease (I00_99)

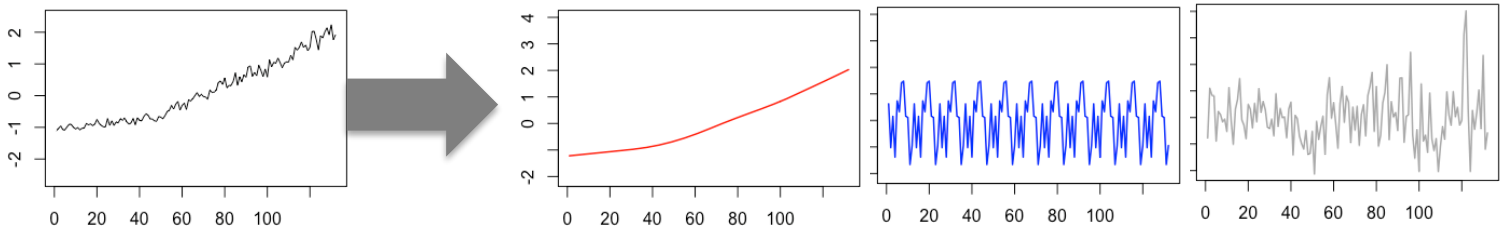
Trend

Seasonality

Remainder



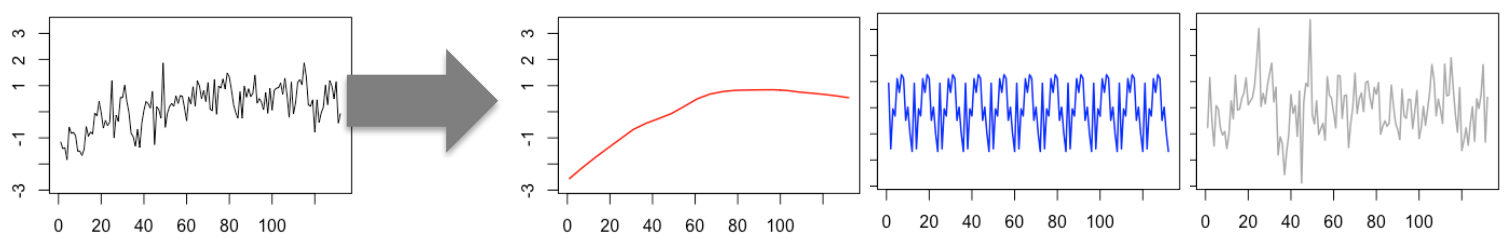
Control disease (A41)



...

...

Control disease (G00_99_SY)



New Approach

Step 1: Decomposition

Step 2: Find the best fitted line

Control disease (I00_99)

Control disease (A41)

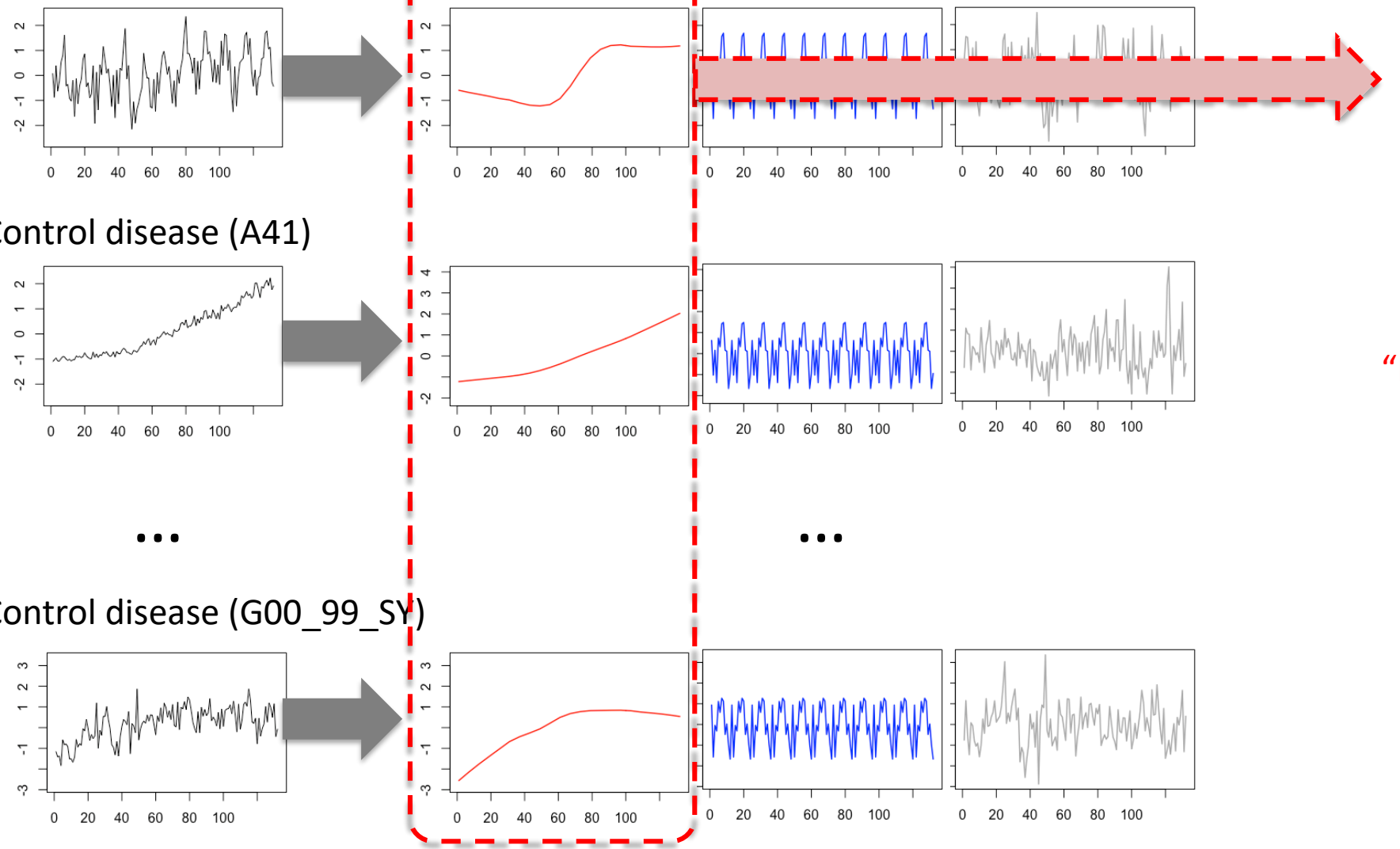
...

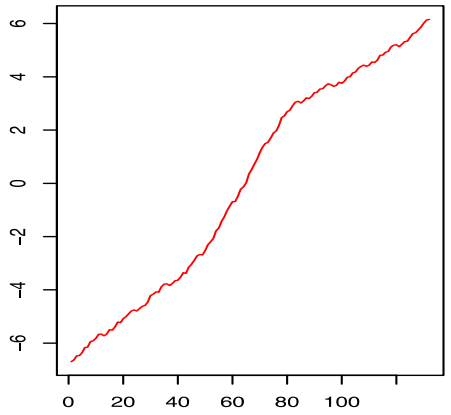
Control disease (G00_99_SY)

Trend

Seasonality

Remainder





"Principal component (PC)"

New Approach

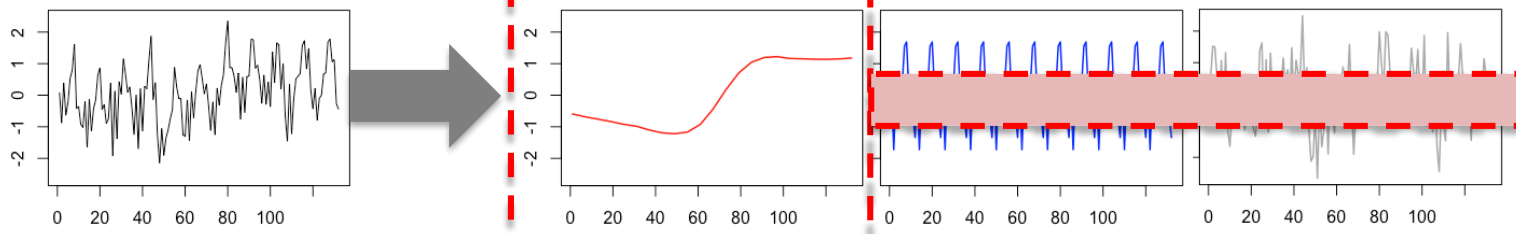
Step 1: Decomposition

Trend

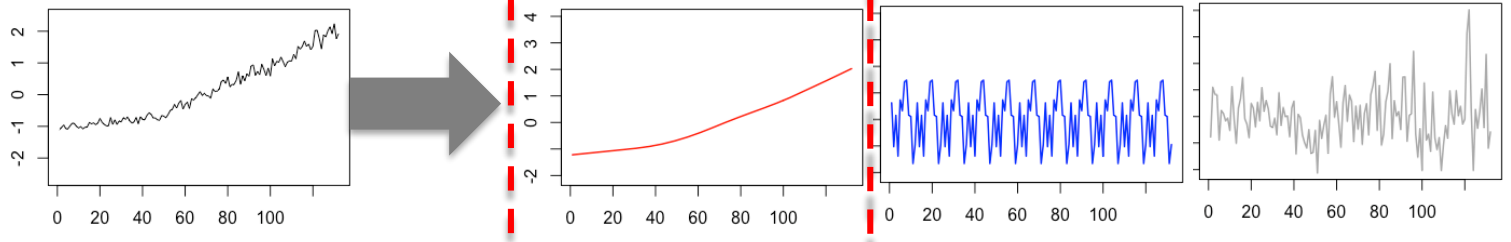
Seasonality

Remainder

Control disease (I00_99)



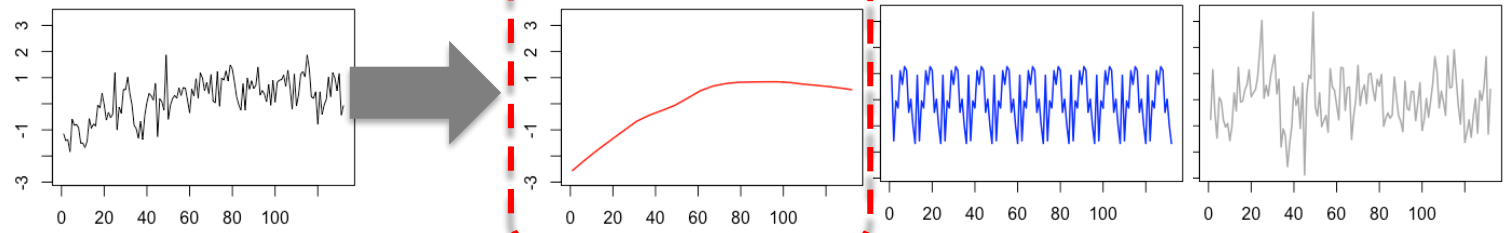
Control disease (A41)



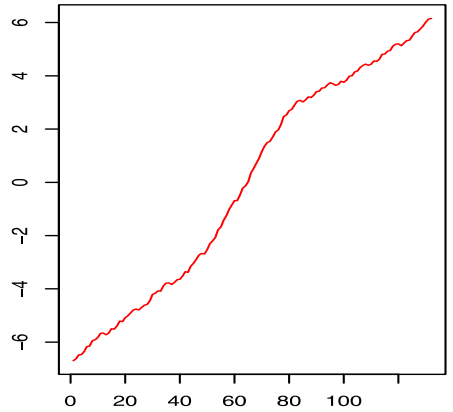
...

...

Control disease (G00_99_SY)



Step 2: Find the best fitted line



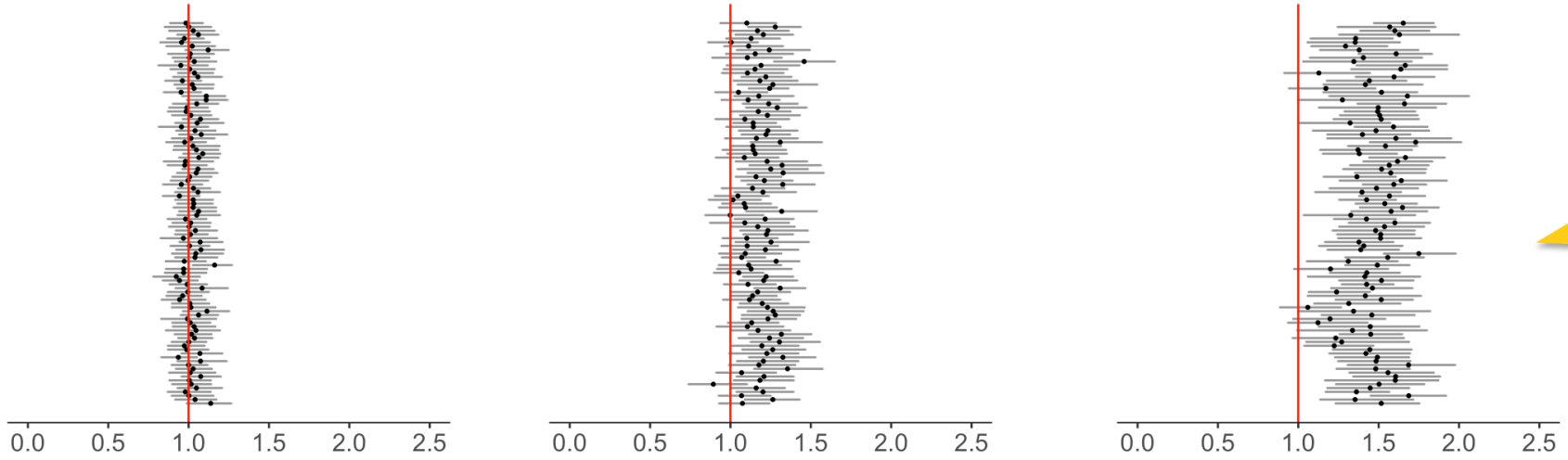
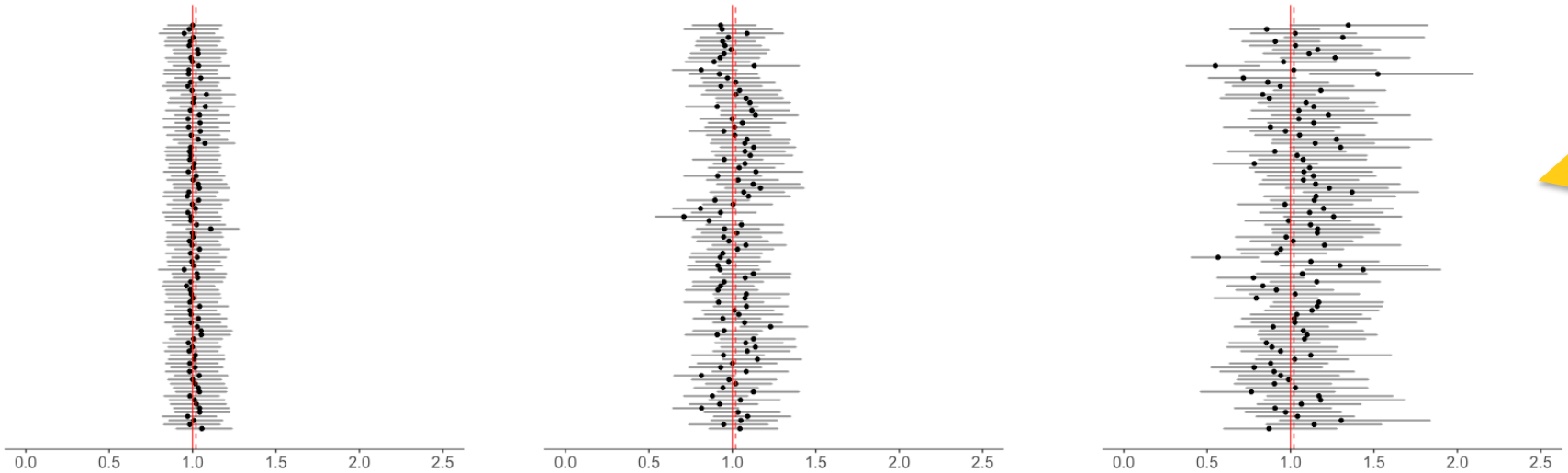
"Principal component (PC)"

Step 3: Regression

$$Pneumonia_t = Intercept + PC + Seasonality$$

New Approach vs. SC model

(National estimate of RR = **0.95** for 80+ yo in Brazil)

	Down samp. rate = 10%	1%	0.25%
SC model			
New approach			

RRs are biased as data become sparse

RRs are not biased. National estimate or RR is included in credible intervals

Summary of New Approach

New approach – 3 steps

1. **Decompose** time series for your control diseases into:
    I. Trend
    II. Seasonality
    III. Remainder
2. Find a **line** that best represents all of the extracted trends
3. Fit **regression** with that line

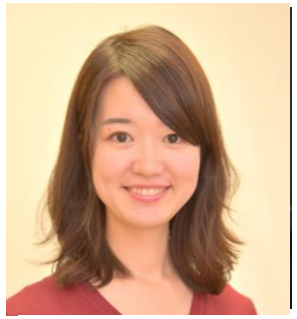
Pros

- Can identify and adjust for unmeasured long-term trends, even when data are sparse and noisy
- Users can simply include all control diseases satisfying the key assumptions in this model
- Regression is very simple

Cons

- Hard to interpret relationships between pneumonia and control diseases, as we are using the best fitted line in the regression

Acknowledgements


Kayoko Shioda


Josh Warren


Lone Simonsen


Christian Bruhn


Cynthia Schuck Paim


Rob Taylor

Yale SCHOOL OF PUBLIC HEALTH

SAGE Analytica
Powerful Insight — Clearly Delivered

BILL & MELINDA GATES foundation

Thank you very much!

Questions?

Dan Weinberger
(daniel.weinberger@yale.edu)

Kayoko Shioda
(kayoko.shioda@yale.edu)