InFER2011
(Inference For Epidemic-related Risk)

Warwick,
28th March-1st April 2011

Statistical inference for models of close-contact infection transmission: An application to varicella in Italy

Luigi Marangi, University of Pisa, Italy.
Emilio Zagheni, MPDIR, Rostock, Germany.
Emanuele Del Fava, University of Hasselt, Belgium.
Ziv Shkedy, University of Hasselt, Belgium.
Gianpaolo Scalia Tomba, UniRoma2, Italy
Piero Manfredi, University of Pisa, Italy.

Outline

- Estimating transmission
  - The problem
  - Traditional approaches
  - Approaches based on social contact data

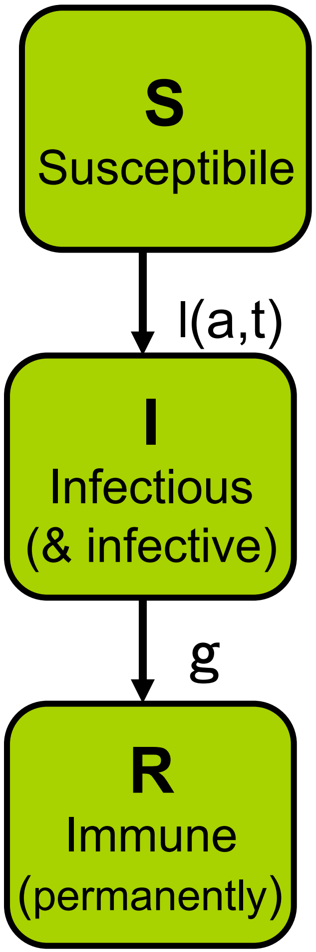
- Inference on transmission parameters
  - A sample of problems: VZV in Italy as case-study.
  - Non-parametric bootstrap inference on transmission parameters.

- Discussion

Dynamic infection transmission models
(infections imparting permanent immunity)

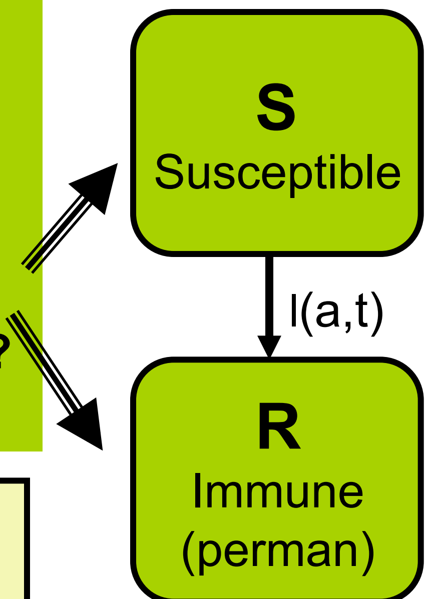
SIR model

l(a,t) = **Force** (“hazard rate”) **of infection**, age–specific
g= force of recovery ➔D=1/g= expected duration infective phase (7 days for VZV)



Disregarding the short infective phase (7days) the FOI “separates” two long phases of host life:
- ❖ the initial susceptible phase
- ❖ the final immunity phase.

Why not to consider an (simpler) SR model ?
(boiling down in a standard survival analysis)



Because evaluation of fundamental parameters (R_0) & policy applications (eg simulating impact of vaccination policies) require using the whole model !

Model equations at equilibrium (eg pre-vaccination period)

$$S'(a) = -\lambda(a)S(a) \qquad S(0) = 1$$
$$I'(a) = \lambda(a)S(a) - \gamma I(a) \qquad I(0) = 0$$
$$R(a) = 1 - S(a) - I(a)$$

S(a) = **Susc fraction aged a**
I(a) = **Infective fraction aged a**
R(a)= **Immune** fraction at age a

Equilibrium FOI :

$$\lambda_i = \sum_{j=1}^{m} q_{ij} C_{ij} \bar{I}_j$$

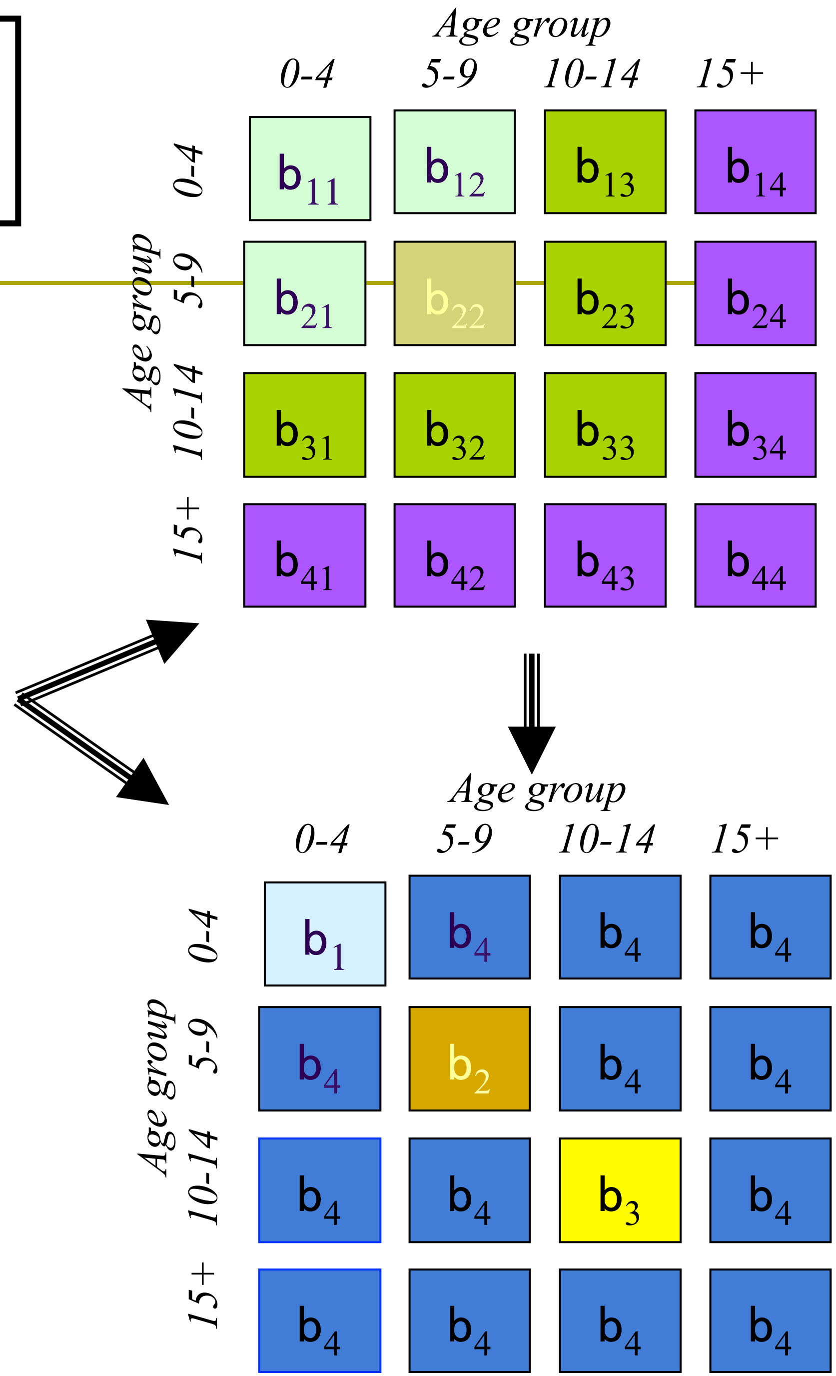
q_{ij}=transmission coefficient per single social contact (age-specific)

C_{ij}= mean number of contacts p.u.t between individuals i and j

2m^2 parameters to be estimated (m=number age groups) !

Traditional approach to estimating transmission

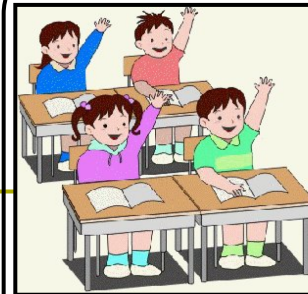
- **Indirect approach** via estimating the FOI from seroprevalence data.
- Use hypotheses ("WAIFW" matrices) to reduce number of unknown parameters from m^2 to m.
- Find transmission rates by solving the (linear) system of equations: $$\lambda_i = \sum_{j=1}^{m} \beta_{ij} \bar{I}_j$$
- **<u>No statistics !</u>**

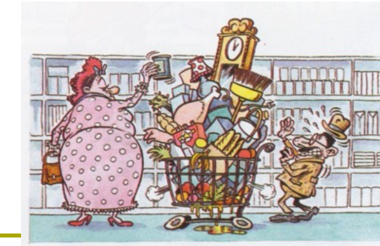
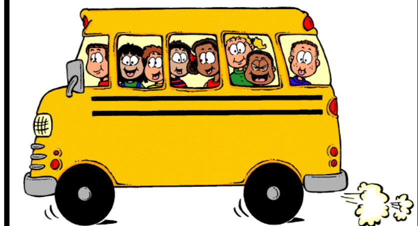


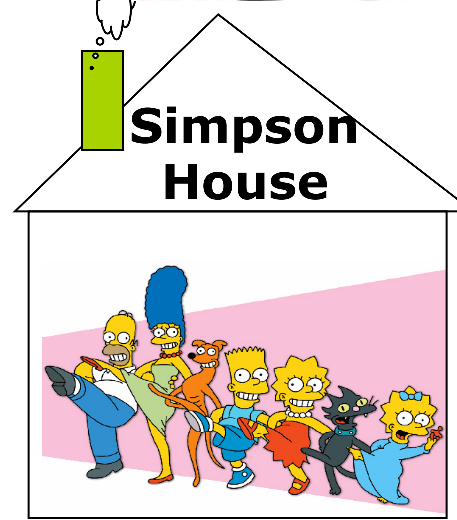
Social contact data

- POLYMOD project (FP6). Direct collection of contact data by contact survey in 8 European countries. (Mossong *et al.* 2008).
- **Definition of "at risk contact"**:
  - "Face to face" conversation.
  - Physical ("skin to skin") contact.
- **Diary-based survey.** Participants reported in a diary all different persons with whom an "at risk contact" occurred in a randomly assigned day. Also reported:
  - age/sex/location of the contact
  - Type (physical/non physical)
  - Duration, etc.

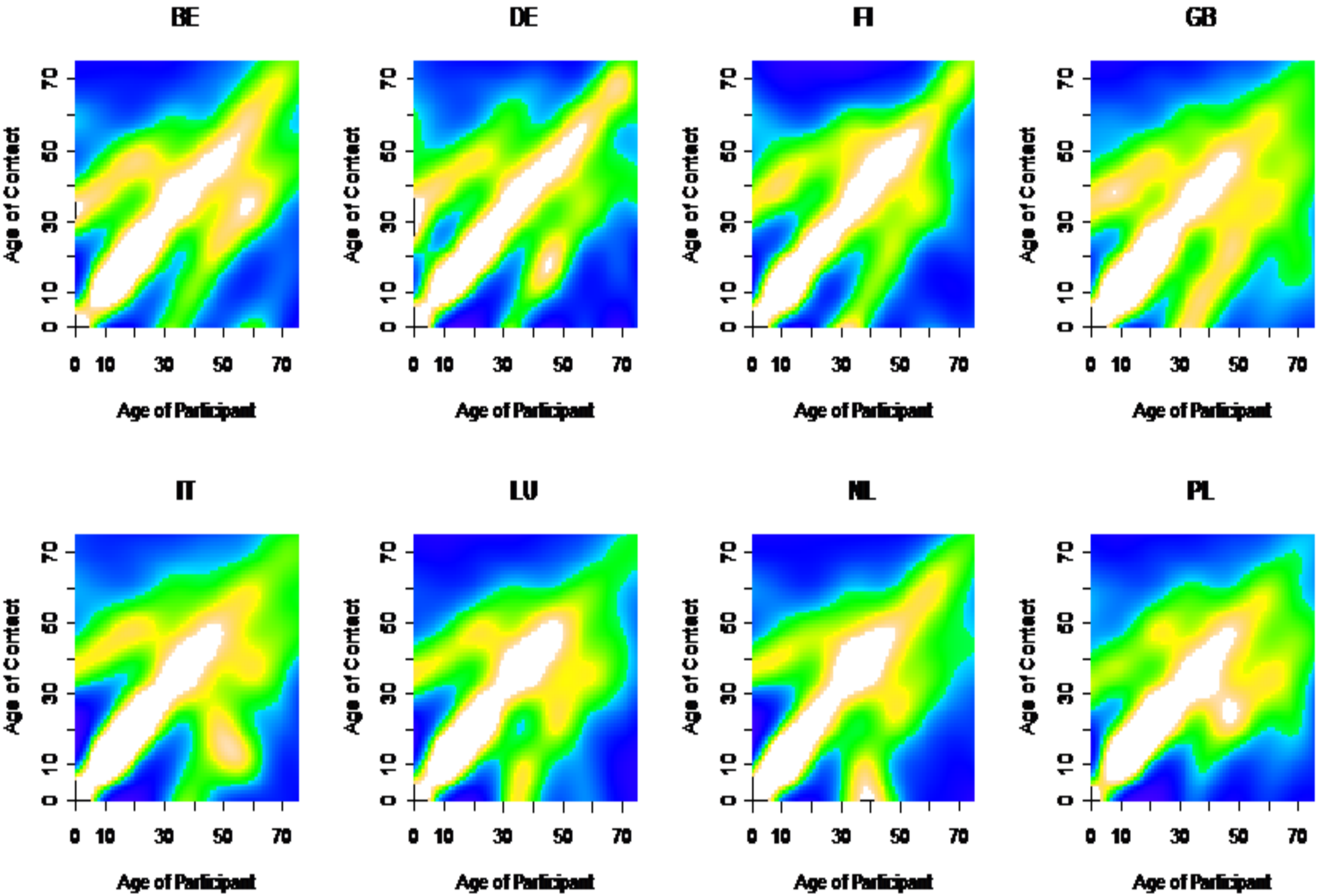
- Also possible to use artificially generated contact data (Del Valle *et al.*, 2007; Iozzi *et al*., 2010).





Source: Mossong et al. 2008



The "social contact" approach to estimating transmission

- Choice of one (or more) contact matrix

- "Social contact hypothesis": reduction of q parameters space by using 1 (constant) "transmission" parameter *q* for each chosen contact matrix

The statistical model: nonlinear regression model linking serological likelihood & contact data

- **Individual immune status**
- **Serological likelihood:**
  - k age groups (k=1…K>m), n_k observations
  - y_k =n. immune individuals
  - π_k =success probability (expected seroprev). It depends on unknown (q_1,..,q_s.) & known pars (C_{ij})
- **(the link)** The expected seroprevalence computed by solving the mathematical model over serological age groups taking contacts as known parameters.

$$Y_i = \begin{cases} 1 & immune \\ 0 & susceptibile \end{cases} \quad i = 1,\ldots,n$$

$$L = L(q_1,\ldots,q_s) = \prod_{k=1}^{K} L_k(q_1,\ldots,q_s)$$

$$L_k = \prod_{i=1}^{n_k} (\pi_k(q_1,\ldots,q_s))^{y_k} (1-\pi_k(q_1,\ldots,q_s))^{n_k - y_k}$$

$$\pi_k = R_k = 1 - \frac{1}{\lambda_k h_k}\left(\prod_{j=1}^{k-1} e^{-\lambda_j h_j}\right)\left(1 - e^{-\lambda_k (a_k - a_{k-1})}\right)$$

$$\lambda_i = \sum_{j=1}^{m} q_{ij} C_{ij} \bar{I}_j = \lambda_i(q_1 \ldots q_s)$$

The estimation problem

Maximise the log-likelihood of the observed seroprofile

$$\log(L) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left[y_k \log \pi_k + (n_k - y_k) \log(1 - \pi_k) \right]$$

$$\pi_k = \pi_k(q_1, \ldots, q_s) = 1 - \frac{1}{\lambda_k h_k} \left(\prod_{j=1}^{k-1} e^{-\lambda_j h_j} \right) \left(1 - e^{-\lambda_k (a_k - a_{k-1})} \right)$$

Subject to:

- The chosen contact matrix & age grouping

$$C = \begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & \ldots \\ \ldots & \ldots & \ldots \end{pmatrix}$$

- $q_i > 0$ (positivity)
- R0>1

$$Dq \|C\| > 1$$

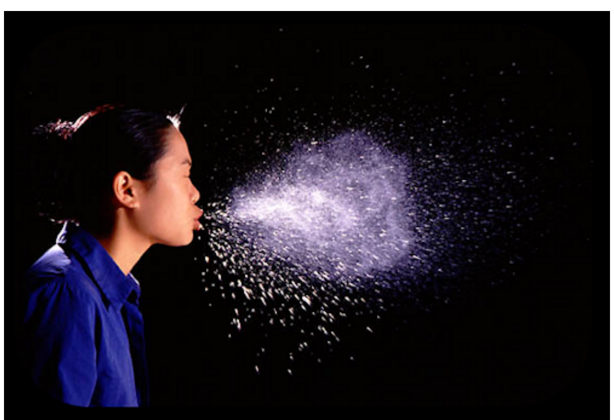
(pre-condition to have an endemic state and therefore to "observe a seroprofile"!)

- FOI equilibrium (discretised integral) equation

$$\lambda_i = \sum_{j=1}^{n} q_{ij} C_{ij} I_j = D \sum_{j=1}^{n} q_{ij} C_{ij} \frac{\prod_{h=1}^{j-1} e^{-\lambda_h h_h} \left(1 - e^{-\lambda_j h_j} \right)}{h_j}$$

Applications: estimating varicella transmission in Italy

- **Etiological agent:** "VZ" virus (herpes virus 3 family (HHV-3).
- **A childhood infectious disease** in industrialised countries.
- **Transmission** via close person-to-person contacts with infective subjects.
- **Duration infectious phase**: about 7 days.
- **Permanent immunity** after recovery.
- However the virus remains latent in the dorsal ganglia, and can reactivate at later ages when immunocompetency declines, causing **herpes zoster** ("shingles").



Experiments

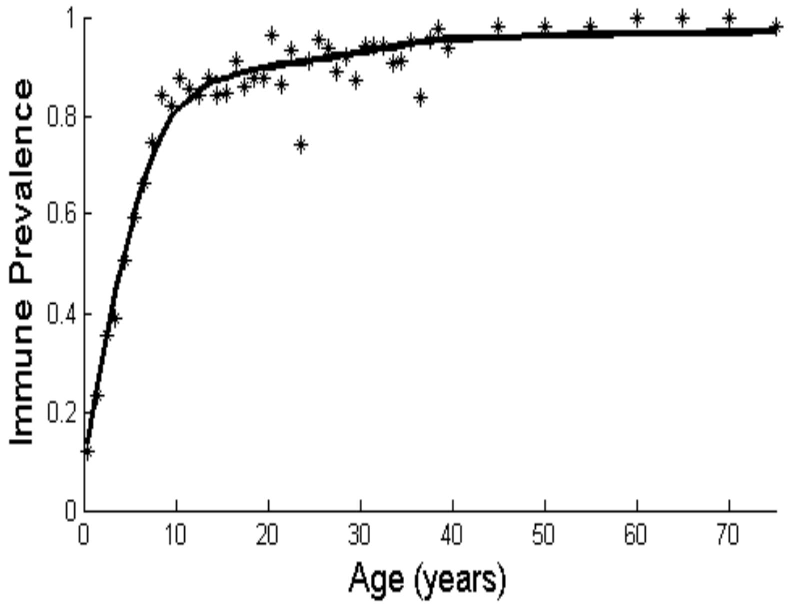
- Estimating simple 1-q models also considered in similar studies (Melegaro *et al.*, *submitted*; Goeyvaerts *et al.*, 2010)
  - M1: "all reported contacts"
  - M2: "physical contacts only"
  - M3: "physical contacts of prolonged duration (>15 min)"

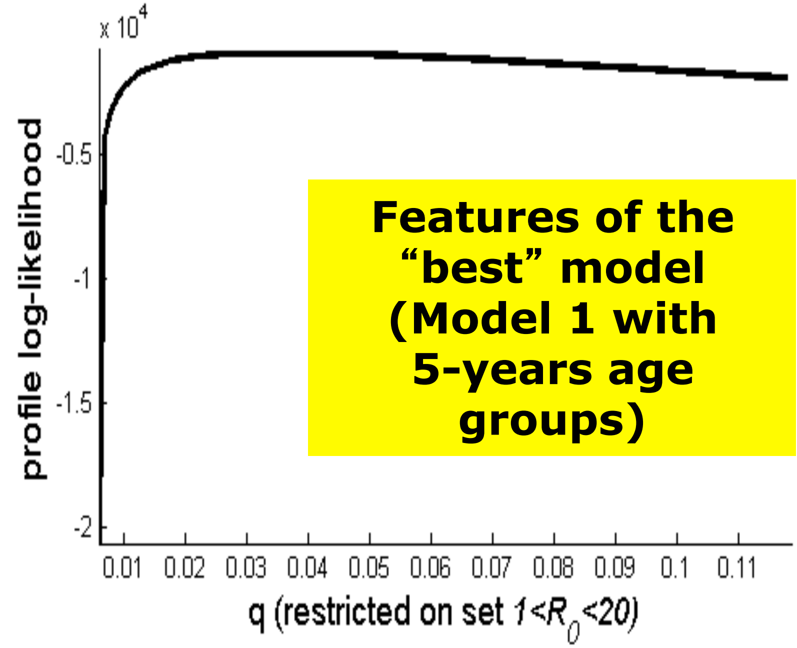
- Bootstrap inference
  - Non-parametric (=re-sampling individuals)
    - serological data
    - contact data
  - "design consistent"
  - Evaluating relative contributions of the two sources of uncertainty
    - separate re-sampling of each source
    - joint re-sampling
  - performances of different types of bootstrap CI (e.g., looking at "real" vs. "nominal" coverage).

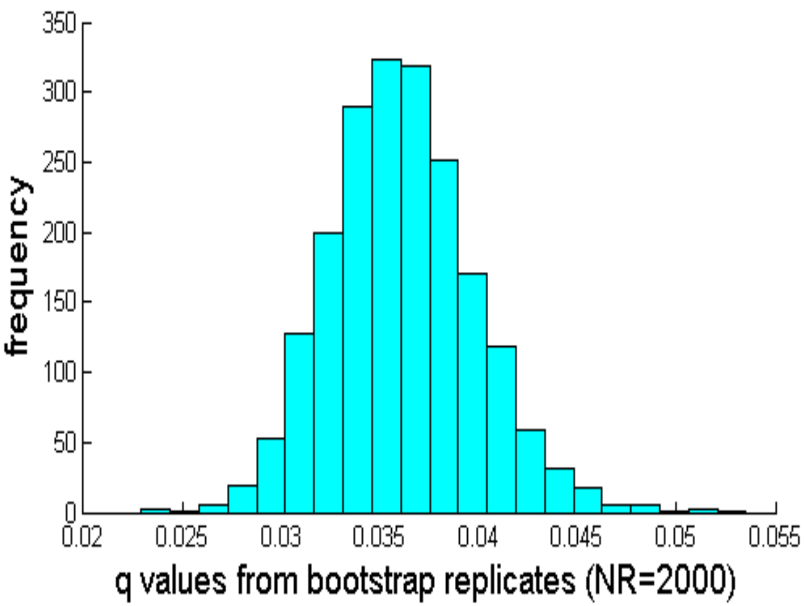
Age grouping	5-years					School	0-2, 3-5, 6-10, 11-13, 14-18, 18-25, etc		
	q	R0	Deviance	AIC		q	R0	Deviance	AIC
Model 1 (1-q)	0,0355	5,97	30,25	1903,10		0,037	6,15	49,38	1922,20
All reported contacts	(0,0339; 0,379)					(0,034;0,0390)			
Model 2 (1-q)	0,049	4,66	32,57	1905,40		0,0510	4,91	60,11	1933,00
Physical contacts only	(0,0465;0,0515)					(0,0484;0,0544)			
Model 3 (1-q)	0,0527	4,18	35,97	1908,80		0,0553	4,49	65,13	1938,00
Physical contacts	(0,0496; 0,0565)					(0,0523;0,0615)			
>15 min									

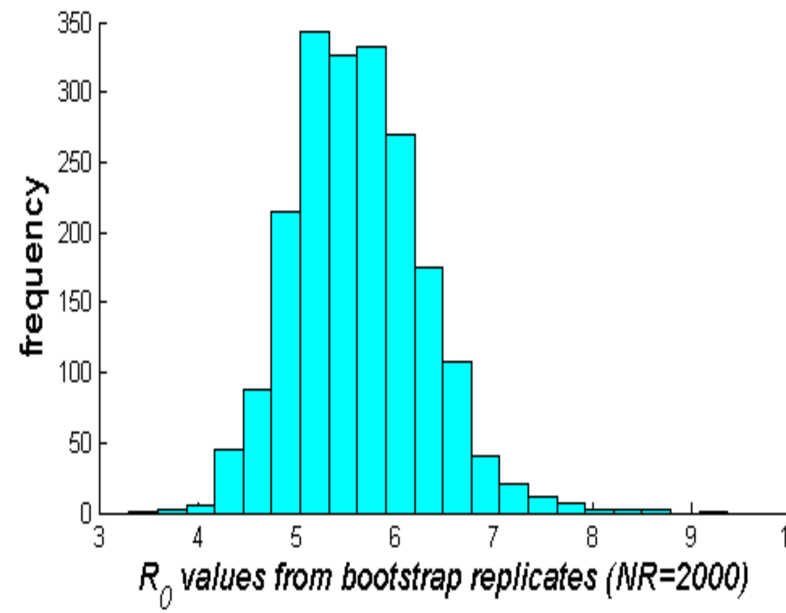
Role of age grouping: 5-years vs school

Features of the "best" model (Model 1 with 5-years age groups)



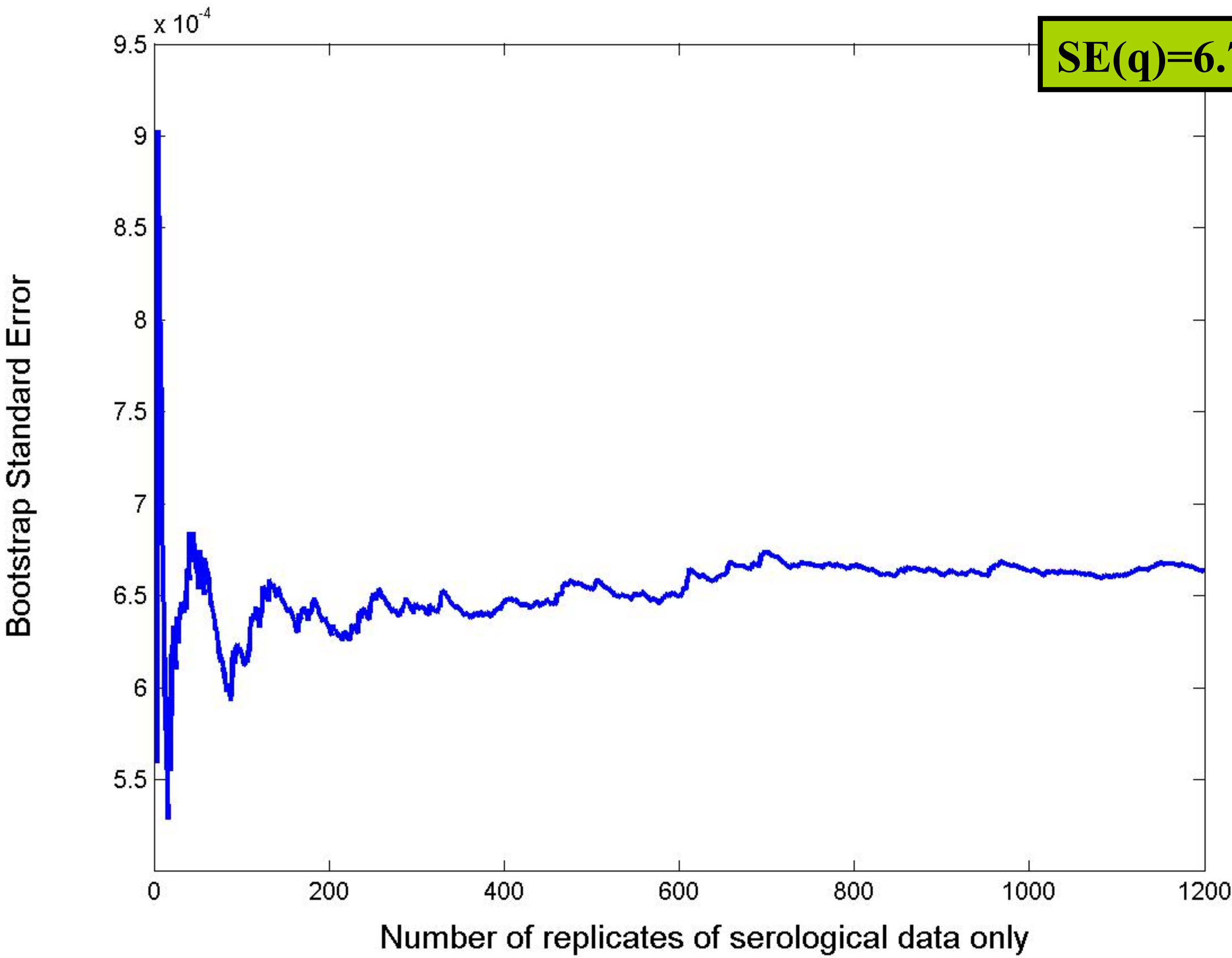






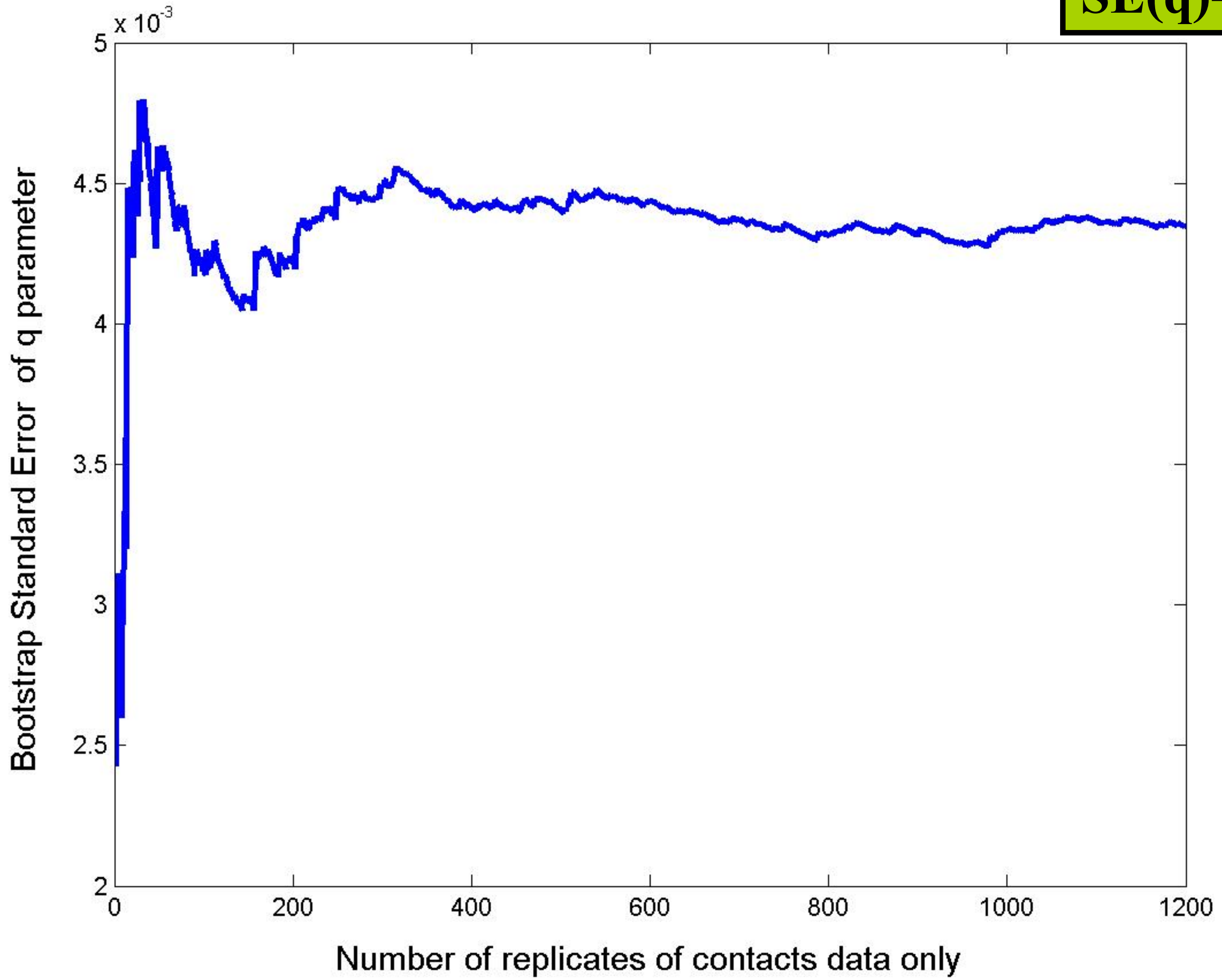
M1 Model, 1-q, “all registered contacts” Bootstrapping (“design-consistent”) serodata only

SE(q)=6.7*10^(-4)



M1 Model, 1-q, “all registered contacts” Bootstrapping (“design-consistent”) contacts data only

SE(q)=4.4*10^(-3)



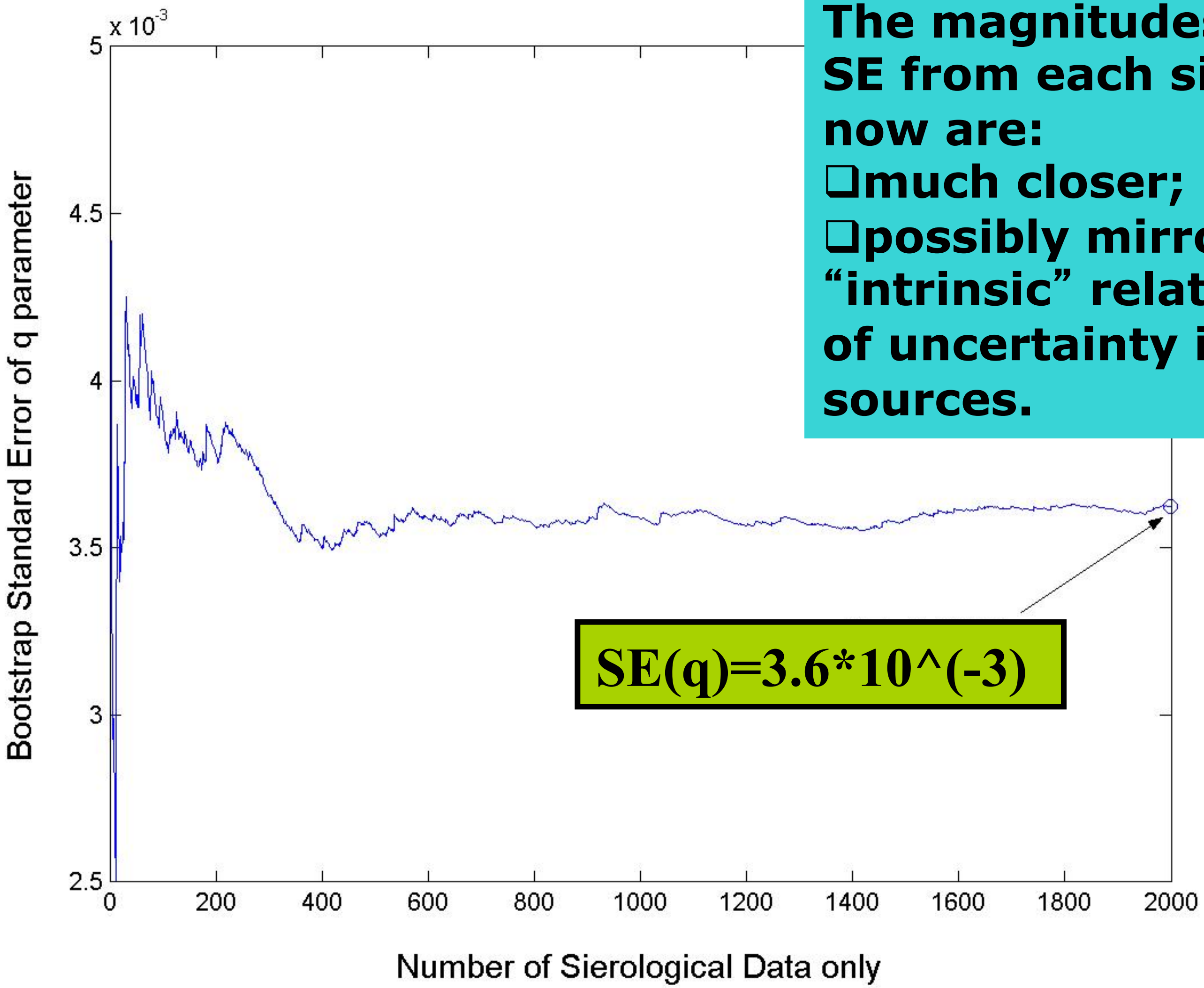
Contacts appear to be the most important source of uncertainty.

The comparison is naive however, given that the two sources concur with largely different numbers of observations: number of serological samples (2446) three times as higher as the number of contact data (845).

Keeping sample size under control:

Keeping fixed the serological proportions per age group, we simulated a serological sample of 845 individuals.

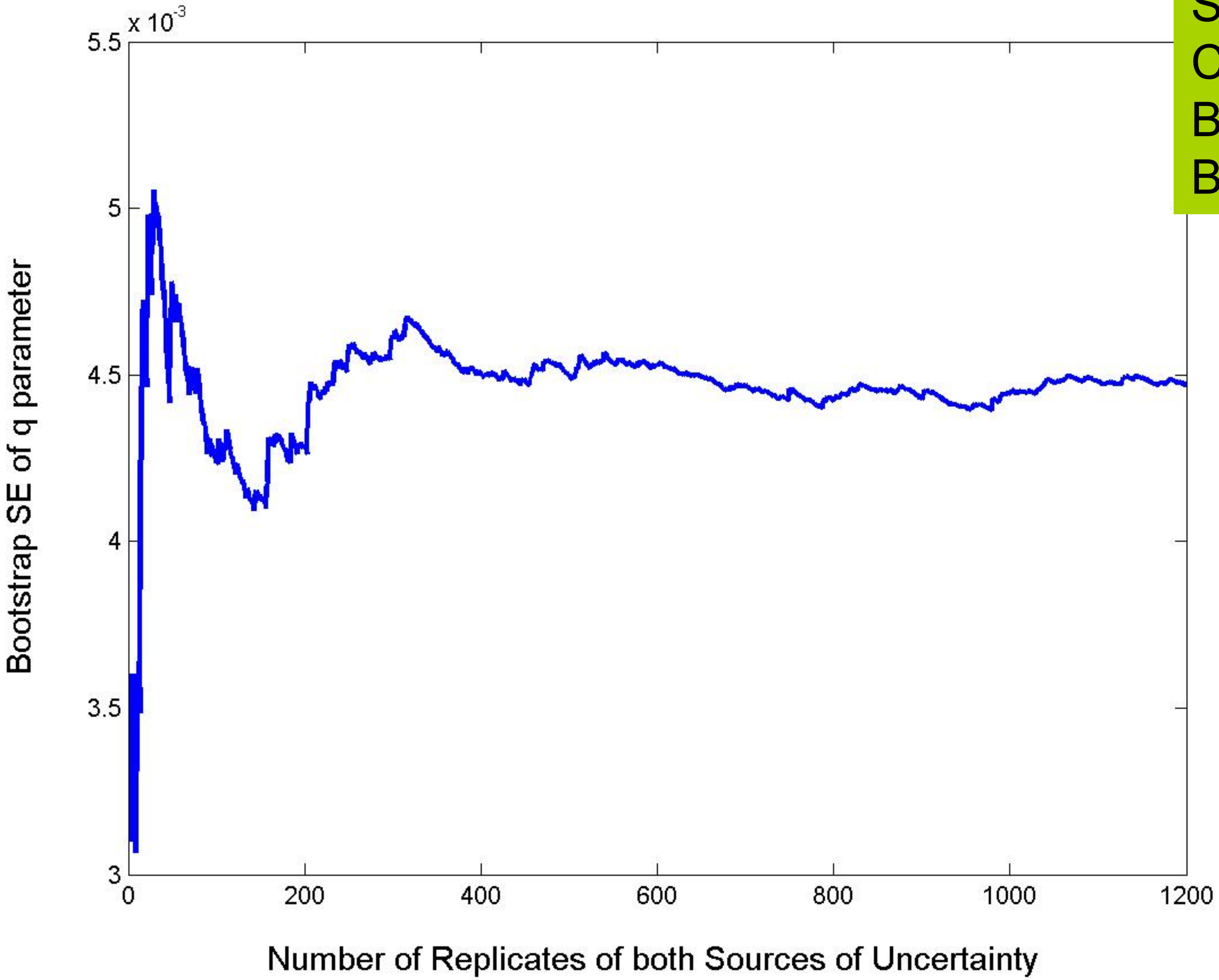
M1 Model, 1-q, "all registered contacts" Bootstrapping sero-data only, but size of serological sample equal to contact sample (other things – e.g., seroprevalence - being equal)



The magnitudes of the "ideal" SE from each single source now are:

- **much closer;**
- **possibly mirroring the true "intrinsic" relative magnitude of uncertainty in the two sources.**

M1 Model, 1-q, "all registered contacts"
Joint bootstrap ("design consistent") of both sources
(1 replicate = 1 resampling from both sources)



SE= 4.5*10^(-3)
CV= 0.1403
BIAS= -1.1*10^(-4)
BIAS/SE= -0.0258

Which bootstrap CI performs better for estimating transmission? (a frequentist experiment: real vs. nominal 95% coverage)

IC	q (Nominal: 1-α=0.95)
Normal	**0.898**
Percentile	**0.937**
BCa (Efron, 1987)	**0.959**

Keeping fixed q, we resampled serological and contact data many times. In this way, we got a series of bootstrap CI's. Then, we computed the proportion of CI's containing the "true" value of q.

Discussion: bootstrap inference & transmission pars

- **Design-consistent bootstrapping per age group provides narrower CI’s with respect to naive resampling (disregarding from the age structure), for any level of nominal coverage.**

- **Contacts appear to be the most important source of uncertainty (under standard sample size).**

- **The bootstrap normal CI shows a real coverage too low, while the BCa CI performs well, as expected.**

Future research

- **Why is a certain contact matrix working better than others? Deepening our understanding of the internal structure of a contact matrix, beyond assortativeness.**

- **Alternative approaches to bootstraap, e.g., MCMC approaches, bayesian melding (Alkema *et al.,* 2007).**

Acknowledgements

- European Centre for Disease Control and Prevention (ECDC) for partial financial support (ECDC grant 2009/002 on Varicella modelling).

- My coauthors, in particular Luigi Marangi who should present this talk (mainly based on Luigi's PhD thesis).

Thank you.