Susceptibles of the world unite! Why we should care about those exposed but not infected when estimating R_0

Eben Kenah
Division of Biostatistics, College of Public Health
The Ohio State University

Joint work with Wasiur Rahman KhudaBukhsh and Grzegorz Rempała

Isaac Newton Institute for Mathematical Sciences

June 23, 2020

Estimation of R_0 without susceptibles

There are two basic methods of estimating R_0 based on generation and serial interval distributions:

- Reconstruct one or more transmission trees and calculate a moving average degree[1]
- Use the exponential growth of the epidemic and a generation interval distribution to calculate R_0 using the Lotka-Euler equation[2]

These are based on branching process approximations to the spread of disease early in an epidemic. They are often simulation tested using branching processes instead of epidemic models with susceptibles in them.

In both approaches, people who were exposed to infection but not infected play no role whatsoever.

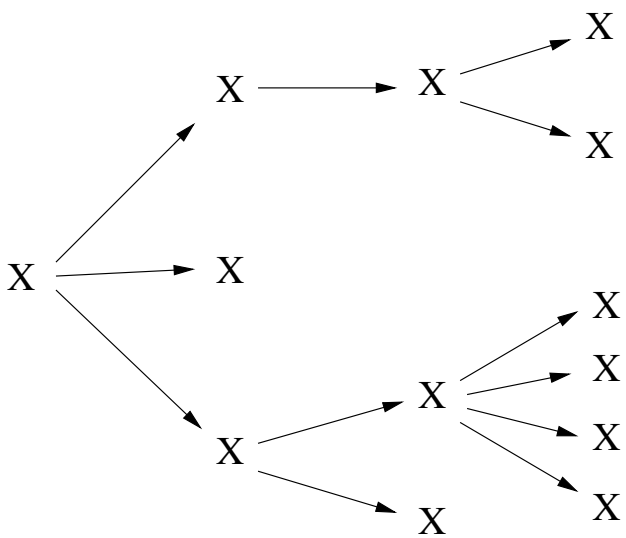
---

[1] J. Wallinga and P. Teunis (2004). *American Journal of Epidemiology* 160(6): 509–516.

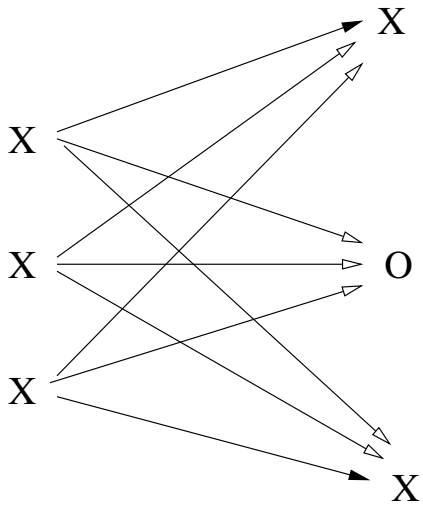
[2] J. Wallinga and M. Lipsitch (2007). *Proceedings of the Royal Society B* 274: 599–604.

Problems shared by both methods

Statistical methods based on generation or serial intervals treat the spread of infection as a branching process, where the generation interval is the time between infections of infectors and infectees.



Branching processes:
People are created when they are infected; susceptibles do not exist, and there is no uninfected person-time.



X = infected person

O = susceptible person

= disease transmission

= failure to transmit

Epidemics:
Infection spreads from infected to susceptible in a preexisting population. Infection is transmitted to a susceptible by the first person to make infectious contact with him or her. Uninfected person-time contains information about disease transmission.

Problems with transmission tree reconstruction

- The probabilities of different trees are generally calculated using a probability density function for the generation interval distribution. These probabilities actually depend on the hazards of infectious contact.
- The mean degree of any tree is just below one, so you are likely to find that your interventions are effective if you follow the epidemic for long enough. Sometimes, additional infections are imputed to avoid this problem.
- The process of forwards and backwards contact tracing to obtain generation intervals and the replacement of generation intervals with serial intervals introduce complex biases.[3]

[3] T. Britton and G. Scalia-Tomba (2019). *Journal of the Royal Society Interface* 16: 20180670

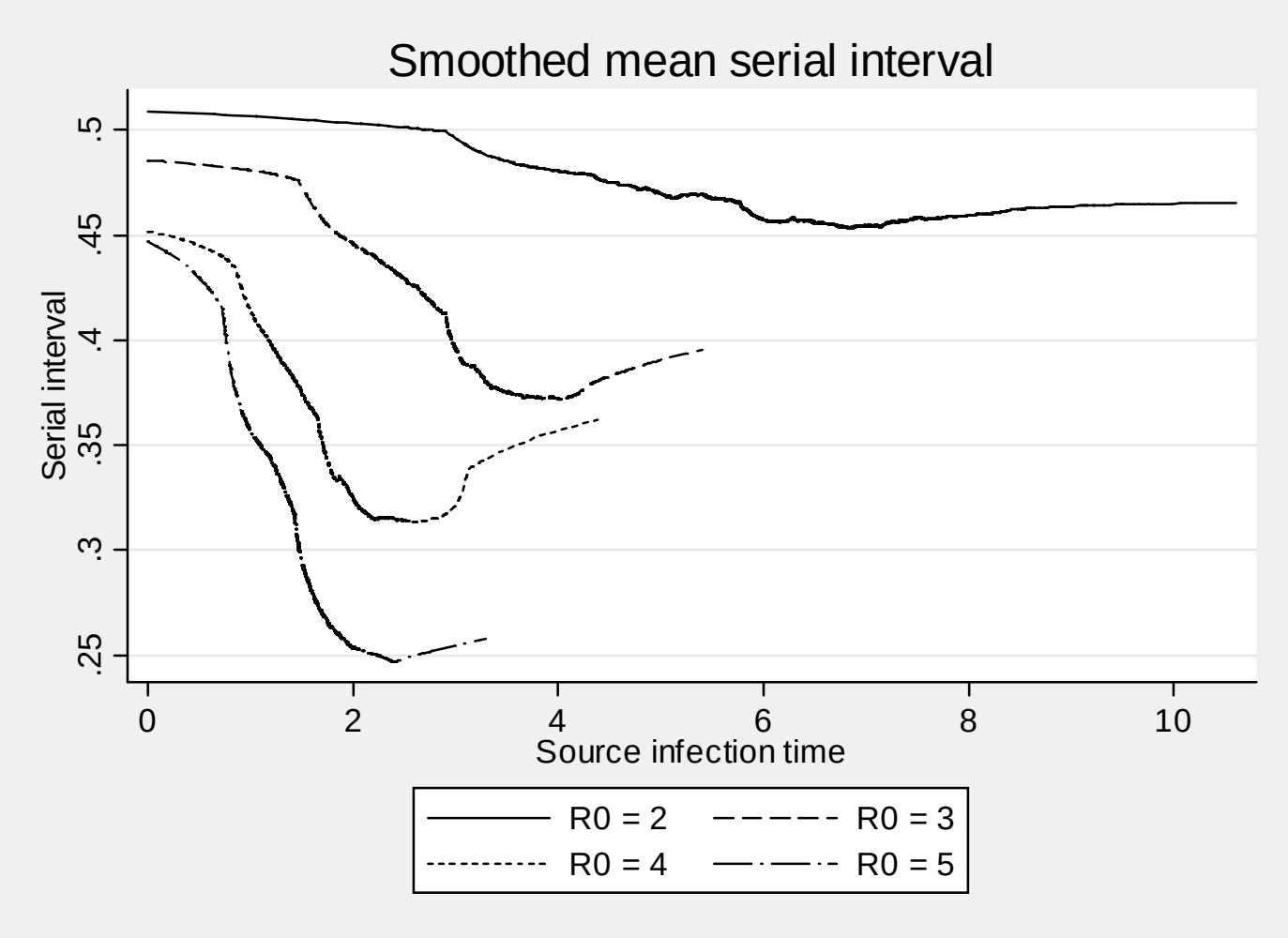
Problems with the Lotka-Euler equation

- These generally assume a mass-action model. Mass-action models have performed poorly in predicting the progress of the epidemic in Ohio and elsewhere.
- The generation interval and serial interval distributions are generally unknown, and they are often estimated from pairs where transmission is considered highly likely. These are unlikely to be representative of all transmissions: Transmission can occur large numbers of low-risk contacts (e.g., public transport, acquaintances, or the grocery store) as well as small numbers of high-risk contacts (e.g., partners and family members).
- The generation interval is not constant. In both mass-action and network-based epidemic models, generation and serial intervals contract due to competing risks of infection and depletion of susceptibles. [4]

[4] Svensson (*Mathematical Biosciences*, 2007) and Kenah, Lipsitch, and Robins (*Mathematical Biosciences*, 2008).

Generation interval contraction

Serial/generation intervals in a mass-action Kermack-McKendrick model

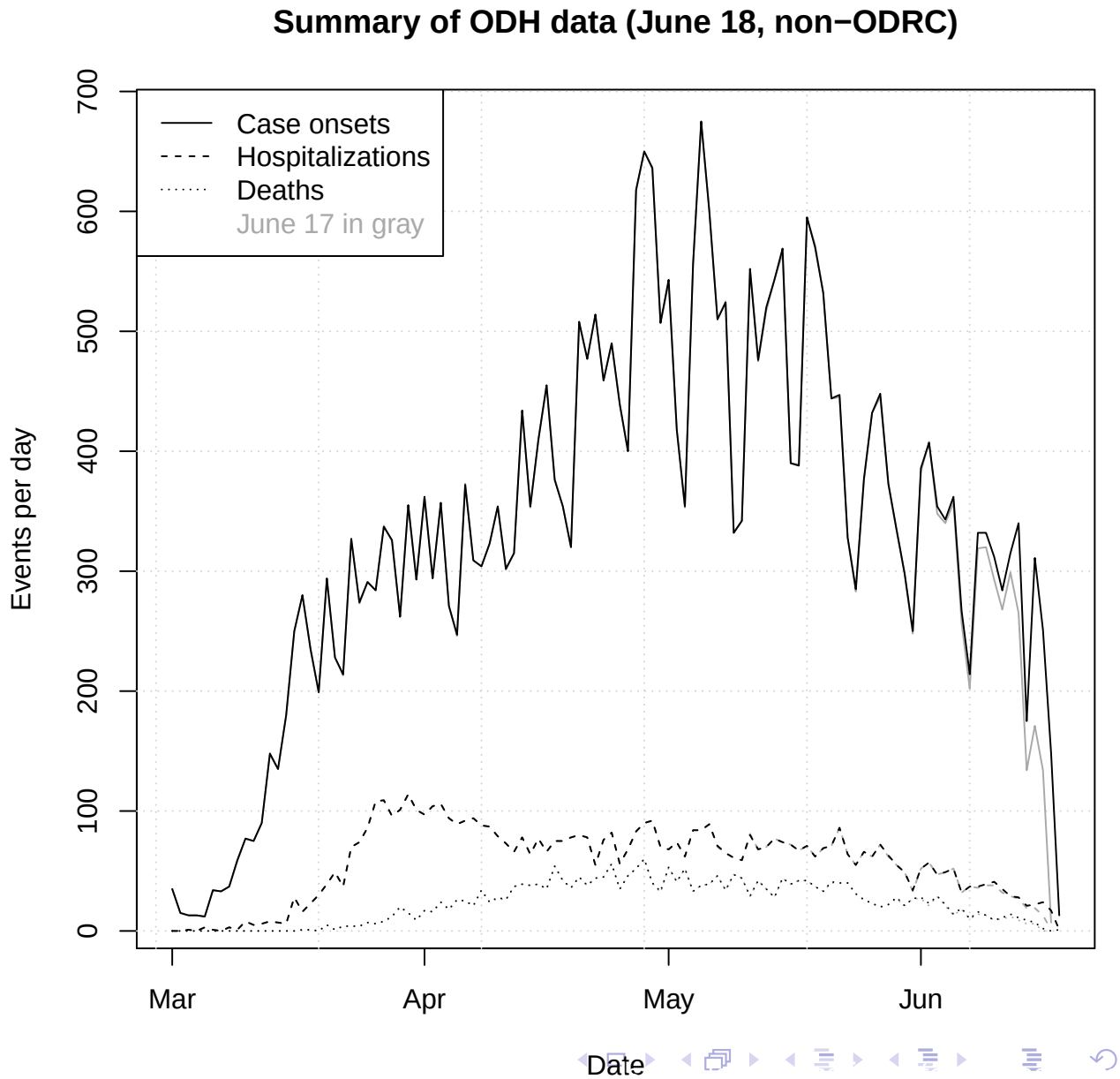


Kenah, Lipsitch, and Robins. *Mathematical Biosciences*, 2008.

COVID-19 dynamics are strange

The COVID-19 epidemic in Ohio and elsewhere has not shown the exponential growth and decay expected of a simple epidemic model.

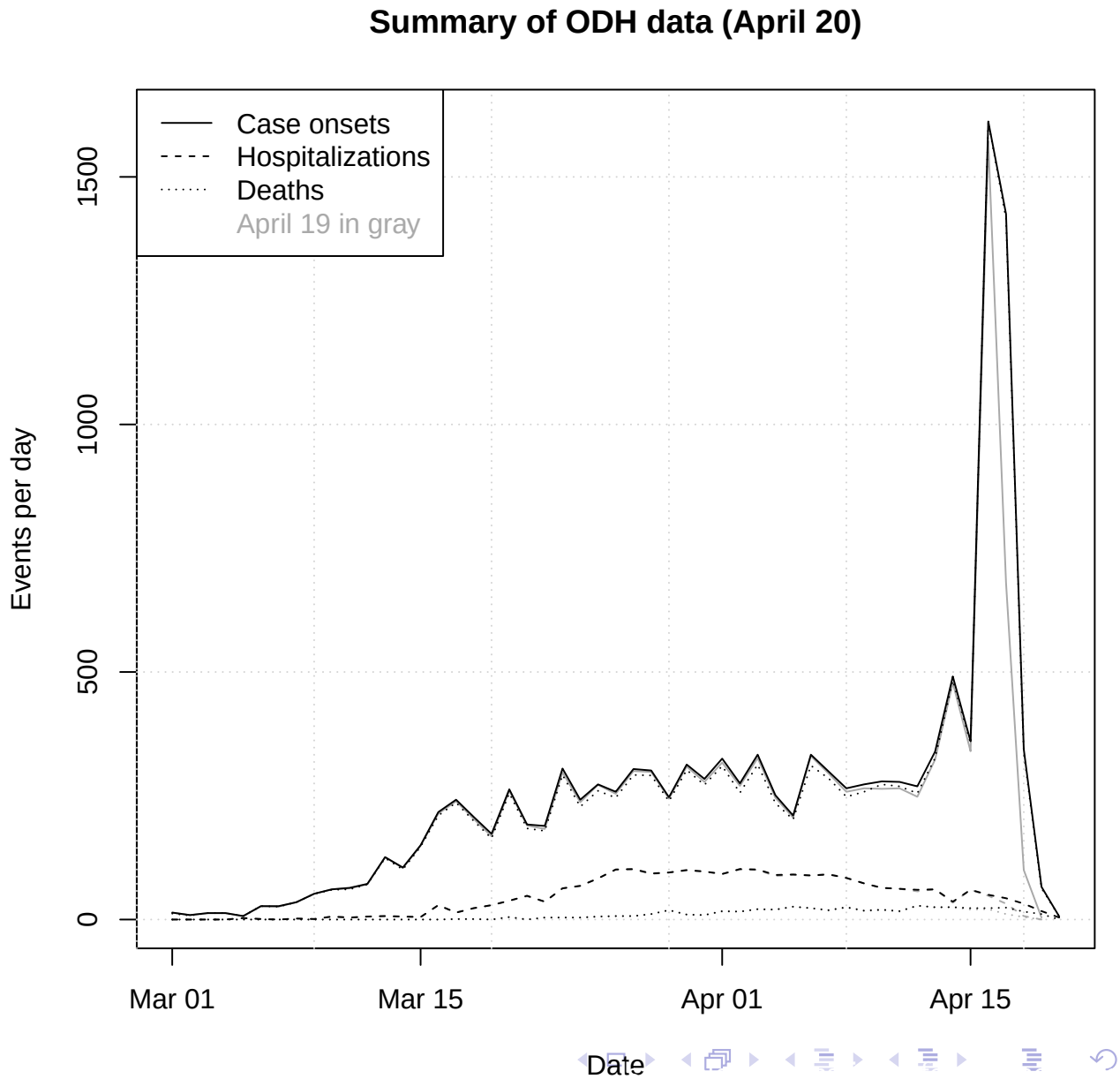
The true epidemic dynamics are obscured by variation in testing supply, demand, and processing.



COVID-19 dynamics are strange

The COVID-19 epidemic in Ohio and elsewhere has not shown the exponential growth and decay expected of a simple epidemic model.

The true epidemic dynamics are obscured by variation in testing supply, demand, and processing.



Contact intervals and pairwise survival analysis

Dependent happenings in infectious disease data can be handled using methods adapted from survival analysis. We define failure times in ordered pairs of individuals, not individuals.

- The *contact interval* τ_{ij} from i to j is the time from the onset of infectiousness in i until infectious contact with j, *whether or not this causes infection in* j.
- Infectious contact is defined to be sufficient to infect j if j is susceptible.
- In an SIR model where i is infected at time t_i, infectious contact from i to j occurs at time $t_{ij} = t_i + \tau_{ij}$. If i is infectious and j is susceptible at time t_{ij}, then i infects j.

Transmission probabilities

The probability of infectious contact from i to j is

$$1 - S_{ij}(\iota) \tag{1}$$

where ι is the infectious period of i and $S_{ij}(\tau) = \Pr(\tau_{ij} > \tau)$ is the survival function of the contact interval distribution.

- If i and j are members of a household, this is the *household secondary attack rate* (SAR).
- Secondary attack rates can also be defined in other settings with a clearly-defined population at risk of infection.

Instantaneous infectiousness

If j is susceptible at time $t_i + \tau$, then j receives infectious contact from i in the interval $(t_i + \tau, t_i + \tau + \mathrm{d}\tau]$ with probability

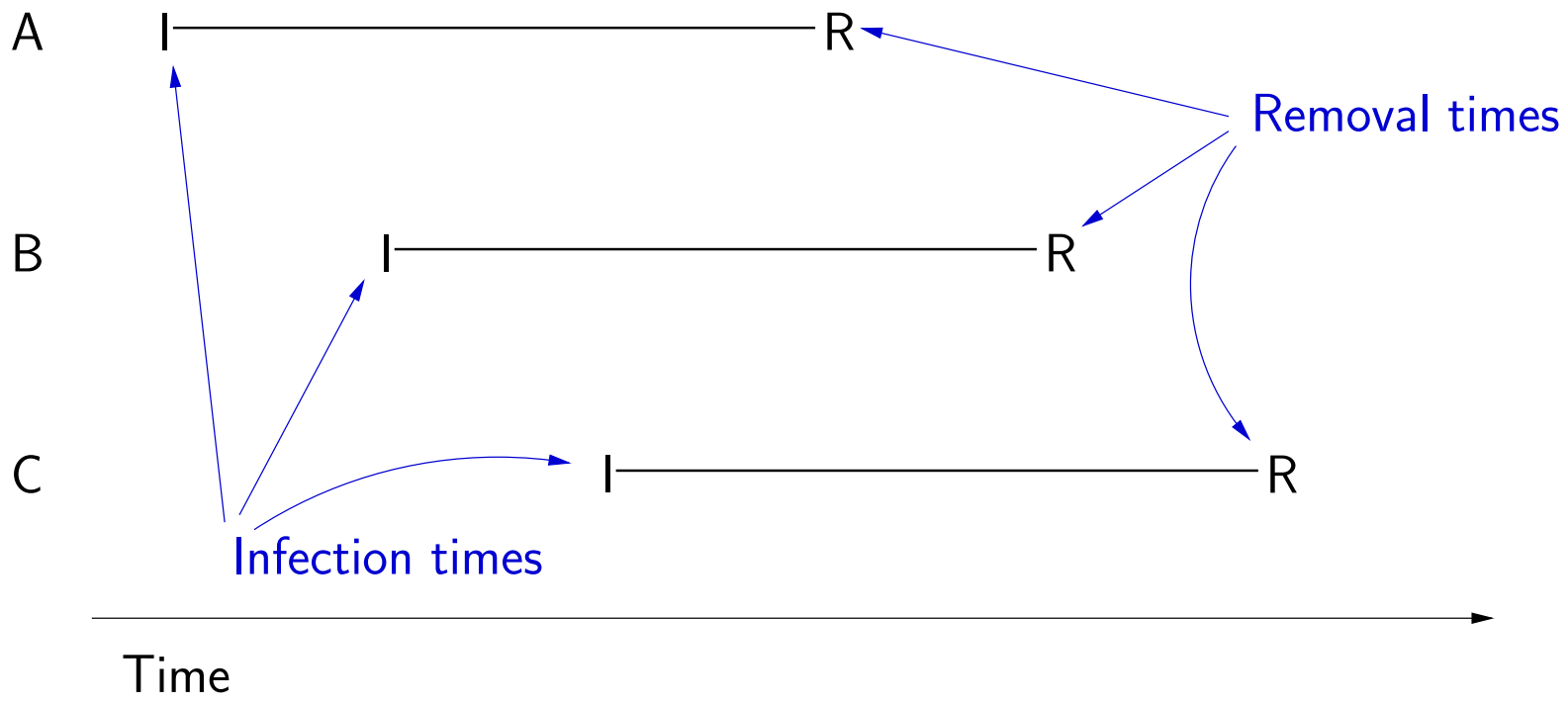
$$h_{ij}(\tau)\,\mathrm{d}\tau \tag{2}$$

where $h_{ij}(\tau)$ is the hazard function of the contact interval distribution.

- The time τ since the onset of infectiousness in i is called the *infectiousness age* of i.
- The function $h_{ij}(\tau)$ gives us the instantaneous infectiousness of i at infectiousness age τ, and it determines the probability of each possible transmission tree.
- A scaled version of this is sometimes called the *infectivity curve* or *infectiousness profile*. It is sometimes scaled to have integral one, sometimes to have integral R_0.

Traditional epidemiologic data

Consider transmission within a household of size three. Members A, B, and C are infected at times t_A, t_B, and t_C respectively. For simplicity, there is no latent period.



Likelihood components

Treating the spread of infection as a branching process, we can write a likelihood in terms of the generation interval distribution PDF.

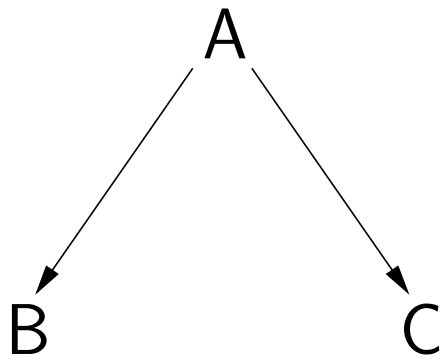
- Let $g(\tau)$ be the PDF of the generation interval.
- Let $g_{ij} = g(t_j - t_i)$.

Treating the spread of infection as a pairwise survival process, we can write a likelihood in terms of the contact interval hazard and survival functions.

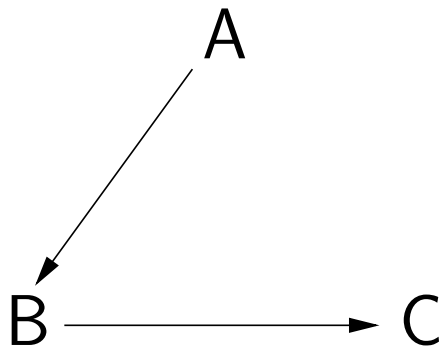
- Let $h(\tau)$ and $S(\tau)$ be the hazard and survival functions.
- Let $h_{ij} = h(t_j - t_i)$ and $S_{ij} = S(t_j - t_i)$.
- The PDF of the contact interval distribution is $h(\tau)S(\tau)$.

Branching process likelihood

There are two possible transmission trees consistent with our data. The likelihood for each tree is the product of generation interval PDFs. The overall likelihood is the sum of the likelihood contributions of each tree.



$g_{AB}g_{AC}$

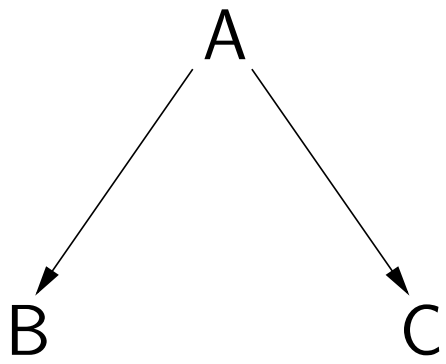


$g_{AB}g_{BC}$

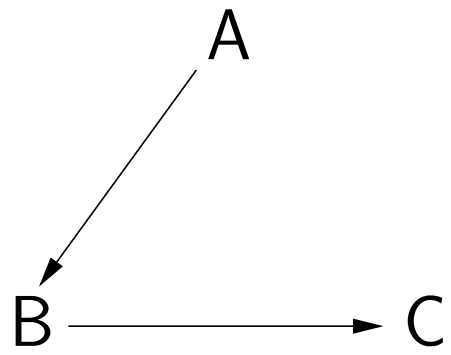
Likelihood: $g_{AB}(g_{AC} + g_{BC})$

Pairwise survival likelihood

The likelihood for each tree is a product of hazard and survival terms. Pairs in which transmission took place get a hazard term. All pairs get survival terms. The overall likelihood is the sum of the likelihood contributions of each tree, which differ only in the hazard terms.



$h_{AB}h_{AC}S_{AB}S_{AC}S_{BC}$



$h_{AB}h_{BC}S_{AB}S_{AC}S_{BC}$

Likelihood: $h_{AB}(h_{AC} + h_{BC})S_{AB}S_{AC}S_{BC}$

Small scales: Pairwise regression

Many methods from standard survival analysis can be adapted to pairwise survival analysis, including

- Parametric estimation (Kenah, *Biostatistics* 2011).
- Nonparametric estimation (Kenah, *JRSSB* 2013).
- Semiparametric relative-risk regression (Kenah, *JASA* 2015).

These are implemented in the `transtat` package for `R`. They can include external sources of infection, and they can be used to simultaneously estimate the effects of individual-level covariates on infectiousness and susceptibility[5]

We plan to use these methods to analyze household data from Guangzhou, China (Jing et al., *The Lancet Infectious Diseases* 2020) and to study transmission of COVID-19 on the Ohio State campus.

[5] Morozova, Cohen, and Crawford (*Journal of the Royal Society Interface*, 2018) showed that estimation of susceptibility alone can be biased across the null even if an exposure is randomized.

Large scales: Dynamic survival analysis

Many mass-action epidemic models and epidemic models on networks can be approximated by systems of ordinary or partial differential equations in the limit of a large population.

Dynamic survival analysis uses these systems of differential equations to define likelihoods for individual infection times and recovery times in a large population. These likelihoods can be seen as large-population limits of pairwise likelihoods.[6]

The fitted model can be used to estimate R_0 and predict epidemic dynamics. This approach has been used by a team of modelers at Ohio State who worked with the Ohio Department of Health to predict hospital and ICU admissions among COVID-19 patients.

[6] W. R. KhudaBukhsh, B. Choi, E. Kenah, and G. Rempała (2019). *Interface Focus* 10: 20190048.

Human sensor networks

Fitting a DSA model does not require daily counts of infections or a known population size. It can be done using a random sample of the population.

Instead of using an epidemic curve to fit a population-level epidemic model, survey sampling techniques could be used to generate a sample of individuals who could be followed longitudinally or panels of individuals at regular intervals.

Both network-based and mass-action DSA models could be fit using this data, and this might provide a practical and accurate method of predicting epidemic dynamics. These individuals could also provide nuclei for studies of transmission in households, workplaces, or other well-defined groups at risk of transmission.

Conclusions

- R_0 is meaningful where there is a reasonable single- or multitype branching process approximation to the spread of disease, but he existence of such an approximation does not imply that a branching process is a good *statistical* model for the estimation of R_0.
- Studies of transmission in well-defined populations at risk of infection have a vital role to play in the estimation of R_0 as well as understanding the natural history of disease, infectiousness, and susceptibility. The lack of such studies of COVID-19 has left many important questions unanswered.
- Testing of random samples of the population (either longitudinal samples or panels) can be used to fit population-level epidemic models that can be used to estimate R_0 and predict epidemic dynamics.
- **Individuals who were exposed to infection but not infected contribute vital information to our understanding of infectious disease transmission.**

Acknowledgements and Funding

The parametric and semiparametric pairwise regression models are joint work with former PhD student Yushuf Sharker, current PhD student Zaynab Diallo, and Wasiur Rahman KhudaBukhsh

This research was supported by the following grants:

- R01 AI116770 *Regression, Phylogenetics, and Study Design in Infectious Disease Epidemiology*
- U54 GM111274 *Center for Inference and Dynamics of Infectious Diseases* (PI: M. Elizabeth Halloran)
- National Science Foundation (NSF) grant DMS 1853587
- WKB is supported by the Presidents Postdoctoral Scholarship Program at The Ohio State University.

The content is solely the responsibility of the author and does not represent the official views of the NSF, NIAID, NIGMS, or the National Institutes of Health.