Stockholms universitet

Introduction to inference for epidemic outbreaks

Tom Britton

September, 2023

Stockholms universitet

Estimation from one large outbreak

Assume a homogeneously mixing community and no preventive measures

From before: in case of a large outbreak and assuming everyone was initially susceptible, the final fraction infected will be close to the positive solution of

$$1 - \tau = e^{-R_0 \tau}$$

Inference other way around: we observe that a fraction $\tilde{\tau}$ got infected. What is R_0?

Rewrite the equation: $R_0 = -\ln(1 - \tau)/\tau$

Our estimate of R_0 is given by the corresponding observed value:

$$\hat{R}_0 = -\ln(1 - \tilde{\tau})/\tilde{\tau}$$

Exercise 14: Estimate R_0 if 20% were infected during an outbreak

Stockholms universitet

Estimation from one large outbreak

This estimate assumed everyone was initially susceptible!

If in fact a fraction r was initially immune we know from before that τ, the fraction *among the initially susceptible* who got infected approximately equals positive solution of

$$1 - \tau = e^{-R_0(1-r)\tau}$$

This leads to the estimate:

$$\hat{R}_0 = -\ln(1-\tilde{\tau})/(1-r)\tilde{\tau}$$

Note: The over all fraction infected equals $\tilde{\tau}(1-r)$

Exercise 15: Suppose as before that 20% were infected during an outbreak, but that only 50% were initially susceptible and the rest were immune. Compute first $\tilde{\tau}$ and then estimate R_0

Estimation of v_c from one large outbreak

It was shown earlier that: $v_c = 1 - 1/R_0$
By observing an outbreak we can hence also estimate v_c (for the same or similar community but not for any community!):

$$\hat{v}_c = 1 - \frac{1}{\hat{R}_0} = 1 - \frac{\tilde{\tau}}{-\ln(1-\tilde{\tau})}$$

If a fraction r was immune in the observed outbreak and $\tilde{\tau}$ of the initially susceptibles were infected this changes to

$$\hat{v}_c = 1 - \frac{1}{\hat{R}_0} = 1 - \frac{(1-r)\tilde{\tau}}{-\ln(1-\tilde{\tau})}$$

Estimation of v_c from one large outbreak

If vaccine not perfect but efficacy E known v_c estimated by

$$\hat{v}_c = \frac{1}{E}\left(1 - \frac{1}{\hat{R}_0}\right) = \frac{1}{E}\left(1 - \frac{(1-r)\tilde{\tau}}{-\ln(1-\tilde{\tau})}\right)$$

Exercise 16. Suppose as previous exercise that 20% of the community got infected but the initial fraction susceptible was 50% (so 40% of these susceptibles were infected). Estimate the critical vaccination coverage for a vaccine having 90% efficacy.

Stockholms universitet

Repetition: Inference from large outbreaks

From before: basic reproduction number R_0 and critical vaccination coverage v_c were estimated by:

$$\hat{R}_0 = -\ln(1-\tilde{\tau})/\tilde{\tau}$$

$$\hat{v}_c = 1 - \frac{\tilde{\tau}}{-\ln(1-\tilde{\tau})}$$

if outbreak takes place in a fully susceptible homogeneous community resulting in a fraction $\tilde{\tau}$ getting infected during the outbreak

How about uncertainty?

Uncertainty of previous estimate

Intuition: The larger community (and more getting infected) the less uncertainty

It was mentioned that final number infected $n\tilde{\tau} = Z$ in case of a major outbreak is normally distributed with mean $n\tau^*$ and standard deviation $\sqrt{n\sigma^2}$ where σ^2 depends on model parameters and shown two slides ahead

This result can be used to show that $\hat{R}_0$ and $\hat{v}_c$ are normally distributed with correct means (i.e. R_0 and v_c respectively) and standard errors to be derived using δ-method

Stockholms universitet

The δ-method

Suppose random variable X has mean $\mu = E(X)$ and variance $V(X)$. Suppose further that we are mainly interested in the distribution of $f(X)$ for some function $f(\cdot)$ rather than X itself

Then the δ-method gives the following approximation for the mean and variance of $f(X)$, where $f(x)$ is a "nice function":

Main idea Taylor expand X around its mean μ:
$f(X) \approx f(\mu) + (X - \mu)f'(\mu)$. This implies:

$$E(f(X)) \approx f(\mu) \qquad V(f(X)) \approx (f'(\mu))^2 V(X).$$

The approximation holds better the smaller variance X has (i.e. smaller $V(X)$).

We will use it for e.g. $f(X) = -\ln(1 - X)/X$ and with $X = \tilde{\tau}$ so that $f(\tilde{\tau}) = \hat{R}_0$

The δ-method for $V(\hat{R}_0)$

Probabilists have proven that the asymptotic variance of $\tilde{\tau}$ equals:

$$V(\tilde{\tau}) \approx \frac{1}{n} \frac{\tau(1-\tau)}{(1-(1-\tau)R_0)^2} \left(1 + c_v^2(1-\tau)R_0^2\right)$$

where τ and R_0 are the true parameter values related by $R_0 = -\ln(1-\tau)/\tau$, and c_v is the coefficient of variation of the infectious period.

We now apply the δ-method on $\hat{R}_0 = -\ln(1-\tilde{\tau})/\tilde{\tau}$, we hence have the function $f(x) = -\ln(1-x)/x$

After some algebra we get $V(\hat{R}_0) \approx \frac{1}{n\tau(1-\tau)} \left(1 + c_v^2(1-\tau)R_0^2\right)$

For a standard error estimate we take square roots and replace unknown quantities with there estimates/observed values. The result, also for $\hat{v}_c$, is given by:

Uncertainty of previous estimate

$$s.e.(\hat{R}_0) = \sqrt{\frac{1 + c_v^2(1-\tilde{\tau})\hat{R}_0^2}{\tilde{\tau}(1-\tilde{\tau})}/n}$$

$$s.e.(\hat{v}_c) = \sqrt{\frac{1 + c_v^2(1-\tilde{\tau})\hat{R}_0^2}{\hat{R}_0^4\tilde{\tau}(1-\tilde{\tau})}/n}$$

$c_v^2 = V(I)/(E(I))^2$= squared coefficient of variation of infectious period of individuals (variance divided by the squared mean)

Larger n gives smaller standard deviation (as expected)!

Uncertainty of previous estimate

c_v^2 cannot be estimated from final outbreak size – possibly known from before

If not one has to insert a "conservative" bound. E.g. $c_v^2 = 1$: very rarely is standard deviation larger than mean

Exercise 25 Suppose that 239 out of 651 individuals in an isolated village were infected during an outbreak. Estimate R_0 and v_c and give 95% confidence interval for the estimates. Consider both the case when all individuals have the same length of infectious period (so no variation) and the case where its standard deviation is equal to the mean.

Exercise 26 Do the same thing assuming 2390 out of 6510 got infected.

Stockholms universitet

Estimation in the early phase of an epidemic

The initial growth: During the early phase of an epidemic incidence as well as prevalence typcially grows exponentially:

$$I(t) \sim e^{rt}$$

ρ (or r) called the **Malthusian parameter**
ρ depends both on R_0 and the generation time distribution $g_0(s)$
Branching process theory: ρ solution to Euler-Lotka equation

$$R_0 \int_0^\infty e^{-\rho s} g_0(s) ds = 1$$

So if we know the generation time distribution $g_0(\cdot)$ we can estimate R_0 from observing the exponential growth ρ!

It is easy to show that if $g_0(s) \sim \Gamma(\alpha, \beta)$ then Euler-Lotka gives that

$$R_0 = \left(\frac{\rho}{\beta} + 1 \right)^\alpha$$

Stockholms universitet

Covid-19: R_0 estimates, **first wave** (original strain)

Covid-19: A common estimate is that $g_0(s) \sim \Gamma$ with mean 6.5 days and s.d. 4 days (see however below!). We assume this to apply to all countries!

We estimate "country" specific ρ from reported cumulative case fatalities: starting first day with > 50 cumulative case fatalities (C_1) and two weeks later C_{15} case fatalities: $\hat{\rho} = \ln(C_{15}/C_1))/14$ (Data: Worldometer)

Common dates: first half of March to end of March (before effects of lockdown)

When 50 have died, between 5 000 and 20 000 had been infected so not VERY early in epidemic which is usually atypical and faster (Norway and Denmark: start instead when > 10 have died)

Covid-19: R_0 estimates, cont'd

Country	C_1	C_{15}	$\hat{\rho}$	$\hat{R}_0$	$\hat{h}_C$
"Norway"	12	89	0.14	2.2	54%
"Denmark"	13	161	0.18	2.6	62%
"Sweden"	62	687	0.17	2.5	60%
"Germany"	68	1275	0.21	3.0	67%
"Belgium"	67	1283	0.21	3.0	67%
"UK"	65	2043	0.25	3.5	71%
"Spain"	55	3647	0.30	4.3	77%

(h_C = critical vaccination coverage for herd immunity)

$\Longrightarrow$ There is not one correct R_0 for covid-19!!

Big differences also within countries!
(Sweden starting when > 10 had died gave $\hat{R}_0 = 3.1$)

Stockholms universitet

Problems with estimating $g_0(s)$ and its consequences

Details: see Britton & Scalia Tomba (2019)

How estimate generation time distribution $g_0(s)$?

Answer: **Contact tracing**: For some identified cases, it is traced by whom and when they were infected

This gives some observed generation times $g_1, \ldots, g_k$. This is often only way, but problematic:

- Generation time defined forward in time but contact tracing backward in time. Problematic?
- For some cases a unique infector and infection time is identified, but for some there are several possibilities (and some have none)
- onset of symptoms more common to observe than infection times
- Identified cases are often severe cases. Do mild/asymptomatic cases have same generation times?

Stockholms universitet

Toy example

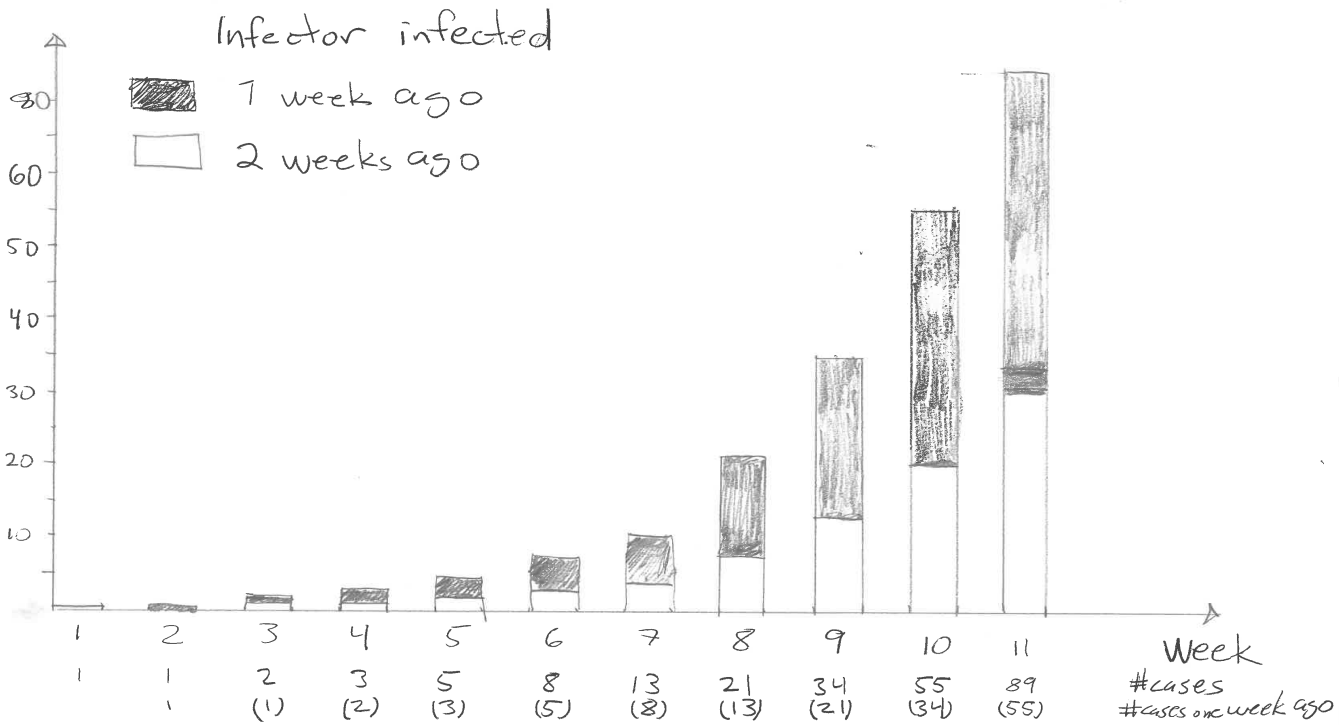
Suppose that $R_0 = 2$, and each infected infects one individual after 1 week and one individual after 2 weeks ($g_0(1) = g_0(2) = 0.5$)

What is $E(G)$?

Toy example

Suppose that $R_0 = 2$, and each infected infects one individual after 1 week and one individual after 2 weeks ($g_0(1) = g_0(2) = 0.5$)

What is $E(G)$? 1.5 weeks, and $st.d.(G)$? 0.5 weeks (below plot of # infections each week)



Looking backwards: contact tracing

Fibonacci numbers and the Golden ratio ...

$\Longrightarrow$ The mean generation time when contact tracing will be < 1.5

So if you estimate $E(G)$ (or all of G) from contact tracing you will *under-estimate* $E(G)$

Stockholms universitet

Generation times vs Serial intervals

Serial intervals instead of generation times

(We now forgetproblem of looking backwards)

Infection times are hardly ever observed, but onset of symptoms are

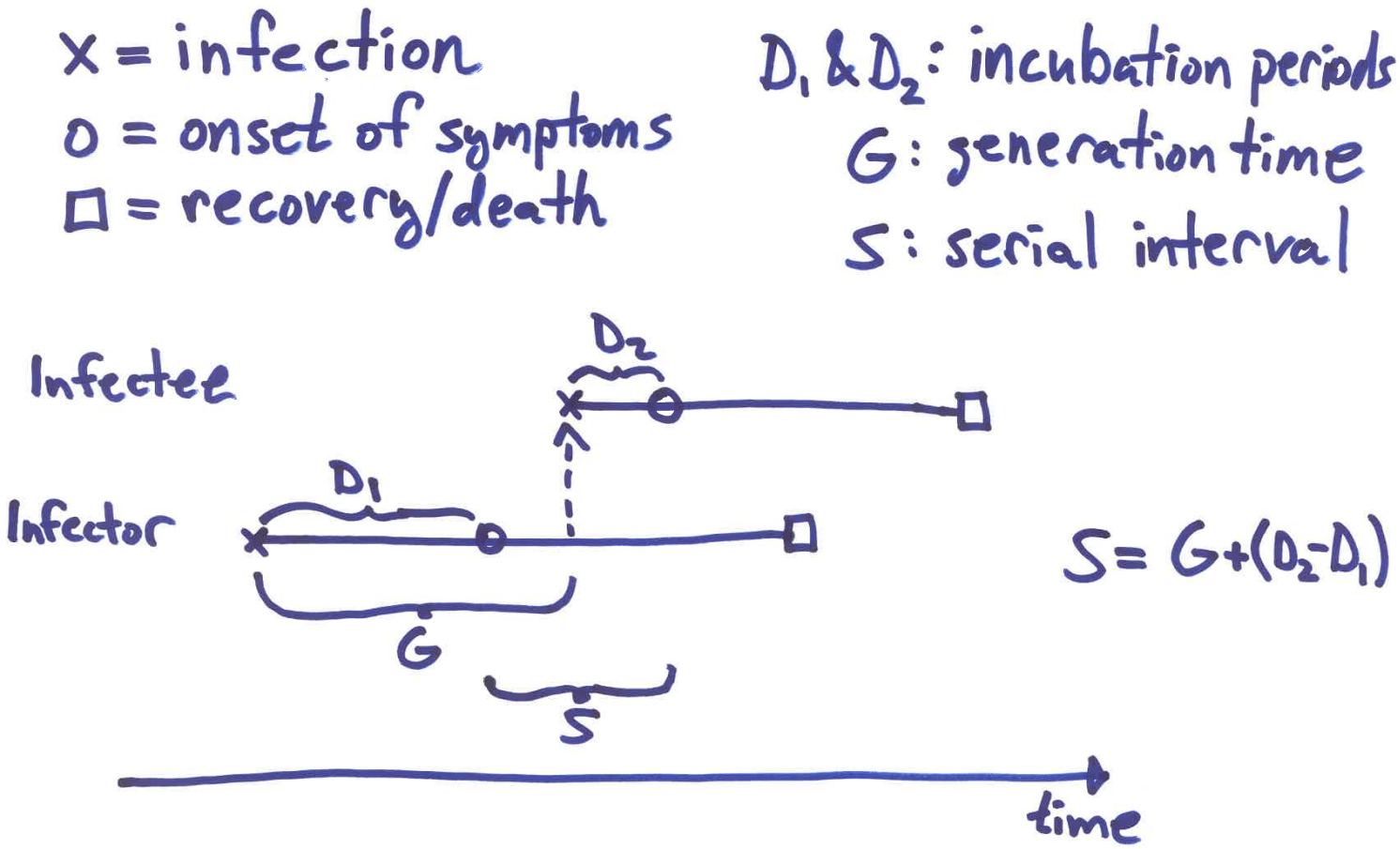
G = time between infection times (unobserved)

S = time between onset of symptoms (observed)

Stockholms universitet

Generation times vs Serial intervals, cont'd

Generaton times vs Serial intervals



Stockholms universitet

Generation times vs Serial intervals, cont'd

$\Longrightarrow S = G + (D_2 - D_1)$ (D_1 and D_2 = incubation periods of infector and infectee)

So, if incubation times are independent and independent of G, then

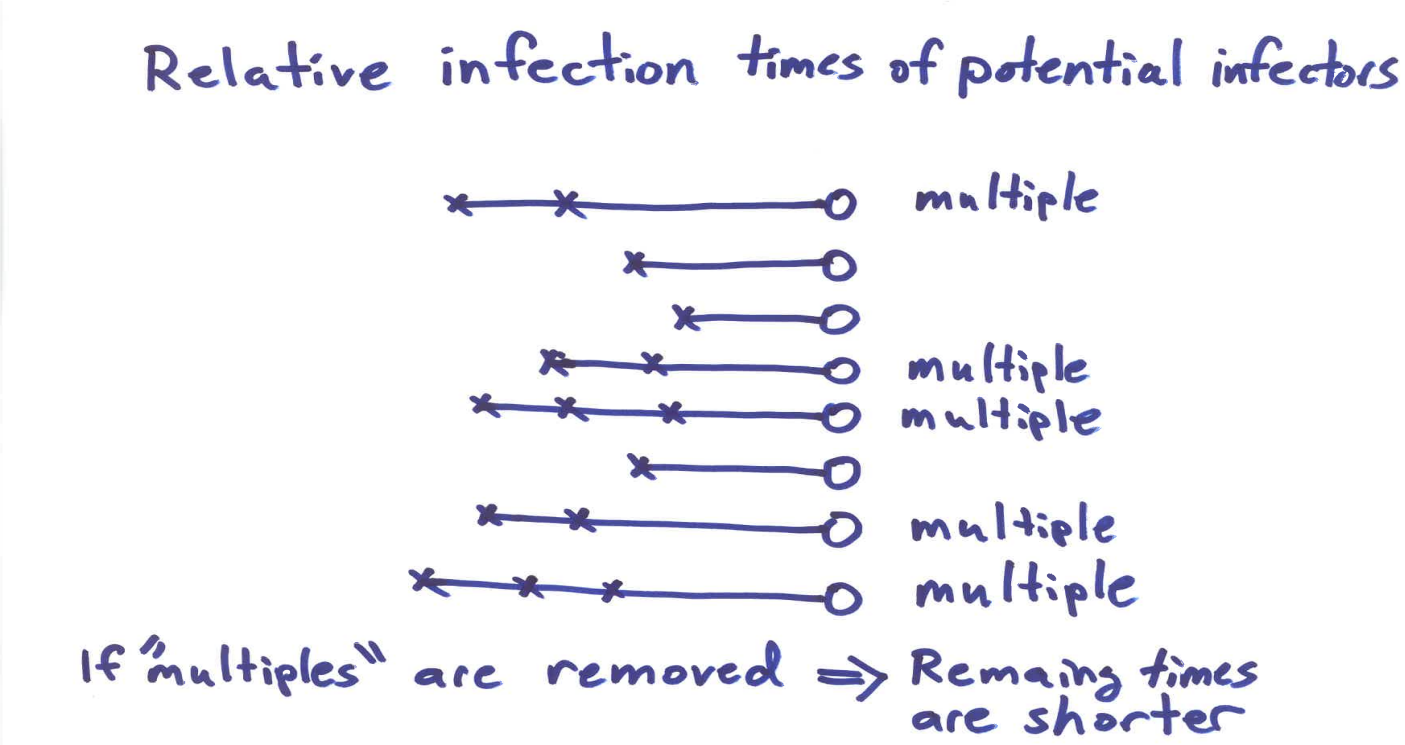
$E(S) = E(G)$, and $V(S) \geq V(G)$

(The relation holds true for all (?) epidemic models)

So, if we estimate $G \sim \{g_0(s)\}$ from observations on Serial intervals we will *over-predict* variance of G

Multiple exposures

Another problem when contact tracing is that sometimes there are several potential infectors (see illustration on next slide)



Multiple exposures

If observations with more than one infected are neglected, remaining intervals are biased from below.

This will also lead to *under-estimation* of $E(G)$

Conclusions: looking backwards and neglecting multiple exposures lead to **under-estimation** of $E(G)$ and observing serial intervals rather than generation intervals lead to **over-estimation** of $V(G)$

We now see how this can affect estimates of R_0

Stockholms universitet

Effects of bias in estimates of $g_0(s)$

$I(t)$ = incidence day t = # infected day t (now discrete time)

How many that get infected day t depends on: R_0 =, basic reproduction number and $\{g_0(s)\}$ = Generation time

– how many that got infected s days ago? Answer: $= I(t-s)$

Model definition (common model)

$$I(t) \sim \text{Pois}\left(R_0 \sum_{s=1}^{t} g_0(s) I(t-s)\right), t = 1, 2 \ldots, \qquad (*)$$

"Pois( )" means Poisson distribution, and the mean equals the parameter, $R_0 \sum_{s=1}^{t} g_0(s) I(t-s)$

Exercise 17.c: Show that this is more or less identical to the Euler-Lotka equation (Hint: replace the Poisson random variable by its mean)

Stockholms universitet

Effects of bias in estimates of $g_0(s)$ (cont'd)

$$I(t) \sim \text{Pois}\left(R_0 \sum_{s=1}^{t} g_0(s) I(t-s)\right), t = 1, 2 \ldots, \qquad (*)$$

If $\{g_0(s)\}$ known (or estimated), Eq. $(*)$ can be used for:

1: Estimating R_0 (from observed incidence $I(1), \ldots, I(t)$), or
2: Predicting outbreak incidence $I(1), \ldots, I(t)$ (if R_0 known before-hand)

Both 1 and 2 require knowledge about $\{g_0(s)\}$

Main question: How to estimate generation time distribution $\{g_0(s)\}$ and what happens to estimates of R_0 (or predictions $I(1), I(2), \ldots$) if $\{g_0(s)\}$ is estimated incorrectly?

Stockholms universitet

Effects of bias in estimates of $g_0(s)$ (cont'd)

Recall, $I(t) \sim \text{Pois}\left(R_0 \sum_{s=1}^{t} g_0(s) I(t-s)\right)$

where $I(0), \ldots, I(t)$ grows, typically exponentially

How are estimates of R_0 (or predictions $I(1), \ldots, I(t)$) affected by the generation time distribution $\{g_0(s)\}$?

Stockholms universitet

Effects of bias in estimates of $g_0(s)$ (cont'd)

Recall, $I(t) \sim \text{Pois}\left(R_0 \sum_{s=1}^{t} g_0(s) I(t-s)\right)$

where $I(0), \ldots, I(t)$ grows, typically exponentially

How are estimates of R_0 (or predictions $I(1), \ldots, I(t)$) affected by the generation time distribution $\{g_0(s)\}$?

It is easy to show that the mean parameter

$R_0 \sum_{s=0}^{t} g_0(s) I(t-s)$ **increases** if:

– $g_0(s)$ is replaced by $\hat{g}_0(s)$ which has smaller mean

– $g_0(s)$ is replaced by $\hat{g}_0(s)$ which has same mean and larger variance

Stockholms universitet

Effects of bias in estimates of $g_0(s)$ (cont'd)

Recall, $I(t) \sim \text{Pois}\left(R_0 \sum_{s=1}^{t} g_0(s) I(t-s)\right)$

where $I(0), \ldots, I(t)$ grows, typically exponentially

How are estimates of R_0 (or predictions $I(1), \ldots, I(t)$) affected by the generation time distribution $\{g_0(s)\}$?

It is easy to show that the mean parameter

$R_0 \sum_{s=0}^{t} g_0(s) I(t-s)$ **increases** if:

– $g_0(s)$ is replaced by $\hat{g}_0(s)$ which has smaller mean

– $g_0(s)$ is replaced by $\hat{g}_0(s)$ which has same mean and larger variance

So, if our estimate of $\{g_0(s)\}$ has mean biased from below we will **under-estimate** R_0

And if we estimate $\{g_0(s)\}$ by something with the correct mean but larger variance we will **under-estimate** R_0

Stockholms universitet

Effects of bias in estimates of $g_0(s)$ (cont'd)

A few slides back we showed three problems when estimating $g_0(s)$ from **contact tracing**:

1) Looking backwards rather than forward in time: $g_0(s)$ was biased from below ($E(G_0)$ under-estimated)
$\Longrightarrow$ R_0 will be **under-estimated**

2) What if multiple infector candidates: $g_0(s)$ was biased from below ($E(G_0)$ under-estimated)
$\Longrightarrow$ R_0 will be **under-estimated**

3) Observing Serial intervals instead of Generation times $g_0(s)$ has too large standard deviation ($V(G_0)$ over-estimated)
$\Longrightarrow$ R_0 will be **under-estimated**

Stockholms universitet

Effects of bias in estimates of $g_0(s)$ (cont'd)

A few slides back we showed three problems when estimating $g_0(s)$ from **contact tracing**:

1) Looking backwards rather than forward in time: $g_0(s)$ was biased from below ($E(G_0)$ under-estimated)
$\Longrightarrow$ R_0 will be **under-estimated**

2) What if multiple infector candidates: $g_0(s)$ was biased from below ($E(G_0)$ under-estimated)
$\Longrightarrow$ R_0 will be **under-estimated**

3) Observing Serial intervals instead of Generation times $g_0(s)$ has too large standard deviation ($V(G_0)$ over-estimated)
$\Longrightarrow$ R_0 will be **under-estimated**

Conclusion: Unless taken account for, all three problems make R_0 *under-estimated*. See Britton & Scalia-Tomba (Interface, 2019)

Biases for Ebola and COVID-19

For Ebola 75% of contacts had multiple potential infectors. The combinded under-estimation of R_0 was $\approx 23\%$

For Corona (Covid19) there was no information of multiple infectors (but I am sure there were!), so only considering bias from backward tracing we believe R_0 is under-estimated by $\approx 12\%$.

Stockholms universitet

The current (or daily) reproduction number R_t

Later on in the epidemic infected individuals no longer infect on average R_0 new individuals for two reasons:

- Some individuals are immune (due to infection and/or vaccination)

- Preventive measures of various forms may have reduced contacts, transmission risks and/or period of infection

If a fraction of immune individuals equals i and the overall reduction in infectious contacts by all preventive measures equals p, then the current reproduction number R_t around calendar time t equals

$$R_t = R_0(1 - i)(1 - p)$$

Stockholms universitet

Estimating R_t from recent incidence

However, if we observe incidence around time t and know the current generation time distribution $g_t(\cdot)$, we can estimate R_t directly from

$$I(t) \sim \text{Pois}\left(R_t \sum_s g_t(s) I(t-s)\right), \qquad (**)$$

(averaging over a few days around t).

But usually $g_t(\cdot)$ replaced by $g_0(\cdot)$ (the initial generation time distribution) ...

Stockholms universitet

GTD changes when preventive measures are adopted

Favero, Scalia Tomba and Britton (2022)

During covid-19 pandemic preventive measure have been enforced and we have changed behaviour:

1. Social distancing in general
2. Self-isolation upon symptoms
3. Screening - testing
4. Contact tracing diagnosed cases

All of these reduce the daily reproduction number R_t (the average number of infections made by an infected now)

But some also change the timing when infections happen, so changes the GTD

Stockholms universitet

A model to investigate effect of prevention on GTD

Contact process:

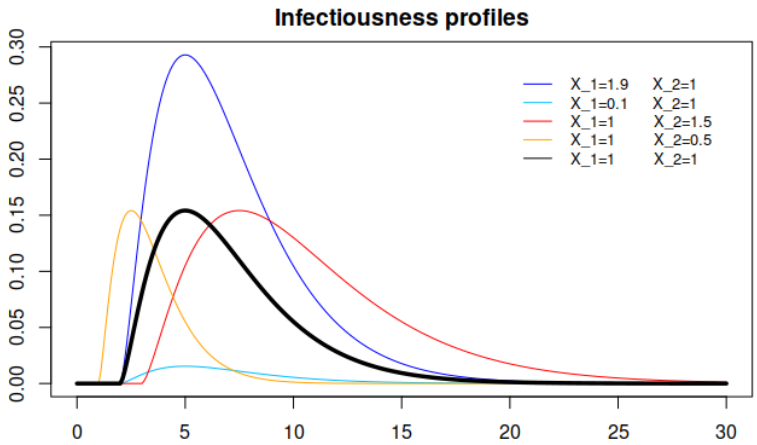
$$C = \{C(t)\}_{t\geq 0} \text{ with } C(t) = \begin{cases} C_1, & \text{if } t \leq \tau \\ C_2, & \text{if } t > \tau \end{cases}$$

C_1: base contact rate (r.v)
C_2: reduced contact rate (r.v)
τ: reduction-time (r.v) e.g. onset or detection

Different definitions of τ, C_1, C_2, allow modelling contacts in several scenarios, with or without interventions

Infectiousness process:

$X = \{X(t)\}_{t\geq 0}$: probability of infection at time t (given a contact)
e.g. $X(t) = p\mathbb{I}_{[0,I]}(t)$ (SIR)
Our focus: $X(t) = X_1 h(X_2 t)$,
h deterministic function, X_1, X_2 r.v.'s



Infectivity proc: $\lambda(t) = C(t)X(t)$

Stockholms universitet

Effects of various preventions:

Infectivity function: $\beta(t) = E(C(t)X(t))$

Basic reproduction number: $R_0 = \int_0^\infty \beta(t)dt$

Generation time density (GTD): $f_G(t) = \beta(t)/R_0$

Various preventions (all reduce R but):

Overall **contact-reduction**: $C \to \rho C$ (no effect on GTD!)

Face masks: $X(\cdot) \to \rho X(\cdot)$ (no effect on GTD!)

Isolation of symptomatic/confirmed: $C_2 \to \rho C_2$ (reduces GTD!)

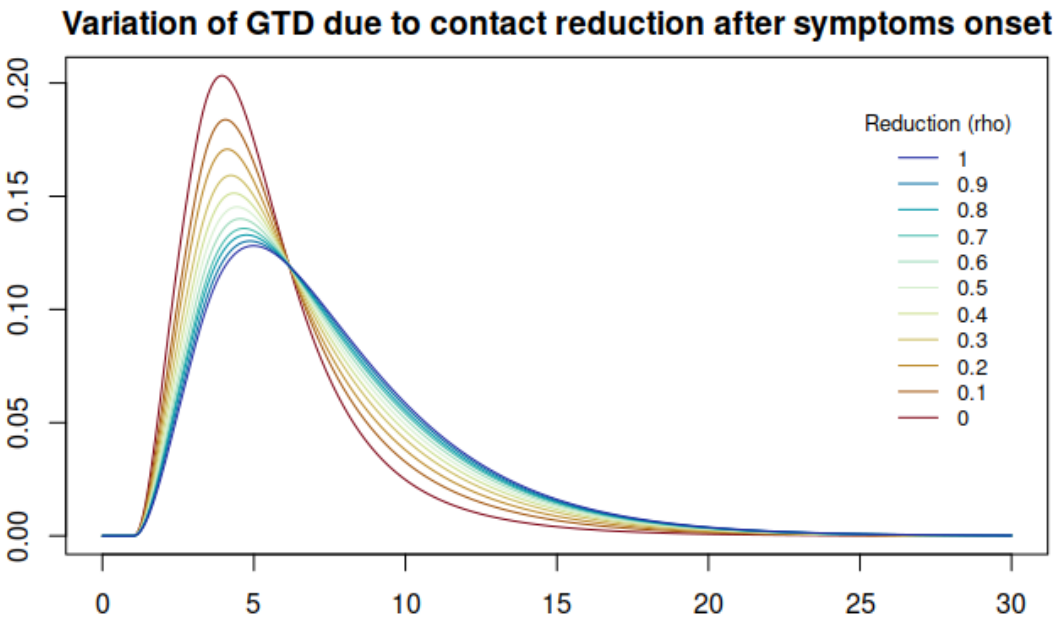
Screening: $\tau = \min\{T_{Sympt}, T_{scre}\}$ (reduces GTD!)

Contact tracing: $\tau = \min\{T_{Sympt}, T_{CT}\}$ (reduces TGD!)

Effects on GTD depends on model assumptions and is quite complicated, in particular contact tracing

Illustration: Isolating symptomatic individuals

$$\tau = T_S \quad C_2 = \rho C_1 \quad X(t) = X_1 h(tX_2)$$



ρ	R	$R^{(1)}$	$R^{(2)}$	mean gen. time (mgt)
1	3.76	1.64	2.11	8.24
0.9	3.54	1.64	1.90	8.11
0.8	3.33	1.64	1.69	7.96
0.7	3.12	1.64	1.48	7.79
0.6	2.91	1.64	1.27	7.60
0.5	2.70	1.64	1.06	7.38
0.4	2.49	1.64	0.84	7.12
0.3	2.39	1.64	0.63	6.81
0.2	2.07	1.64	0.42	6.44
0.1	1.85	1.64	0.21	5.98
0	1.64	1.64	0	5.41

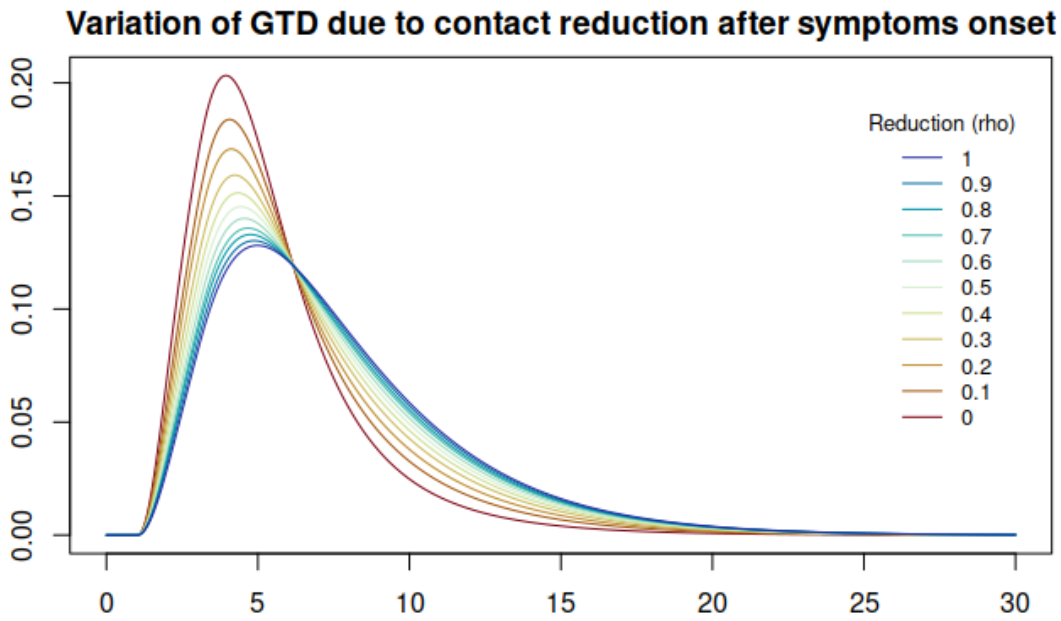
Asymptomatic cases: about 1/3

Example: $\rho : 0.5 \to 0.1$ implies R reduced by 31% and mgt by 19%

Stockholms universitet

Illustration: Isolating symptomatic individuals

$\tau = T_S \quad C_2 = \rho C_1 \quad X(t) = X_1 h(tX_2) \quad$ MGT= mean generation time



ρ	R	$R^{(1)}$	$R^{(2)}$	MGT
1	4.54	1.73	2.81	7.57
0.9	4.26	1.73	2.53	7.48
0.8	3.98	1.73	2.25	7.38
0.7	3.70	1.73	1.97	7.28
0.6	3.42	1.73	1.69	7.15
0.5	3.14	1.73	1.41	6.99
0.4	2.86	1.73	1.13	6.82
0.3	2.57	1.73	0.84	6.59
0.2	2.29	1.73	0.56	6.31
0.1	2.01	1.73	0.28	5.96
0	1.73	1.73	0	5.48

Asymptomatic cases: about 1/3

Example: $\rho : 0.5 \to 0.1$ implies R reduced by 36% and mgt by 15%

Stockholms universitet

Covid example and effect on bias

Combining preventions (added isolation, screening and CT) where we have "guessed" suitable values reduces

$R = 3.9 \ \rightarrow \ R = 1.45$ (reduction by 62%)

$E(G) = 7.4 \ \rightarrow \ E(G) = 5.8$ days (reduction by 22%)

Inferring R_t

Suppose we observe (increasing) incidence $\{I(t)\}$ for this situation ($R_t = 1.45$ and mean gen-time $E(G) = 5.8$)

If we use this new correct GTD and apply Euler-Lotka estimating equations we get $\hat{R}_t \approx 1.45$ as it should

However, if we instead used the original GTD with mean 7.4 days (as most do!) we would get $\hat{R}_t \approx 1.75$, so biased by $> 20\%$

R_t-**estimates** that use early GTD-estimates are **biased from above** (or more accurately "biased away from 1")

Stockholms universitet

Thanks for your attention!

References

- Ball, F., Britton, T., Leung, K. and Sirl, D. (2019). A stochastic SIR epidemic model with preventive dropping of edges. *J. Math. Biol.* **78**:1875-1951

- Ball F and Britton T (2021). Epidemics on networks with preventive rewiring. *Rand Str Alg*. **61**:1-48. https://doi.org/10.1002/rsa.21066

- Britton, T, Ball F, Trapman P. (2020). A mathematical model reveals the influence of population heterogeneity on herd immunityto SARS-CoV2. *Science*. **369** (6505), pp. 846-849. DOI: 10.1126/science.abc6810

- Britton, T., Janson, S., Martin-Löf A. (2007): Graphs with specified degree distributions, simple epidemics and local vacination strategies. *Adv. Appl. Prob.*, **39**: 922-948.

- Britton T and Leskelä L (2022). Optimal intervention strategies for minimizing total incidence during an epidemic. *Submitted*. https://arxiv.org/abs/2202.07780

Britton, T. and Scalia Tomba G. (2019). Estimation in emerging epidemics: biases and remedies. Journal Royal Society: Interface. 16:20180670

- Leung, K., Ball, F., Sirl, D. and Britton, T. (2018). Individual preventive social distancing during an epidemic may have negative population-level outcomes. *Journal Royal Society: Interface*, **15**: 20180296.