1 Introduction

2 Box: Kappa tutorial

3 Results

Demographic stochasticity can generate "emergent" heterogeneity even in the absence of explicit differences between individual-based rates. In simple models, this heterogeneity can be characterized. We explicate the notion that this is predictable (see Box). **[JD:** *Is that really what Box is doing, though? Or more about linking the emergent stochasticity in the deterministic vs. demographic-stochastic models?***]**

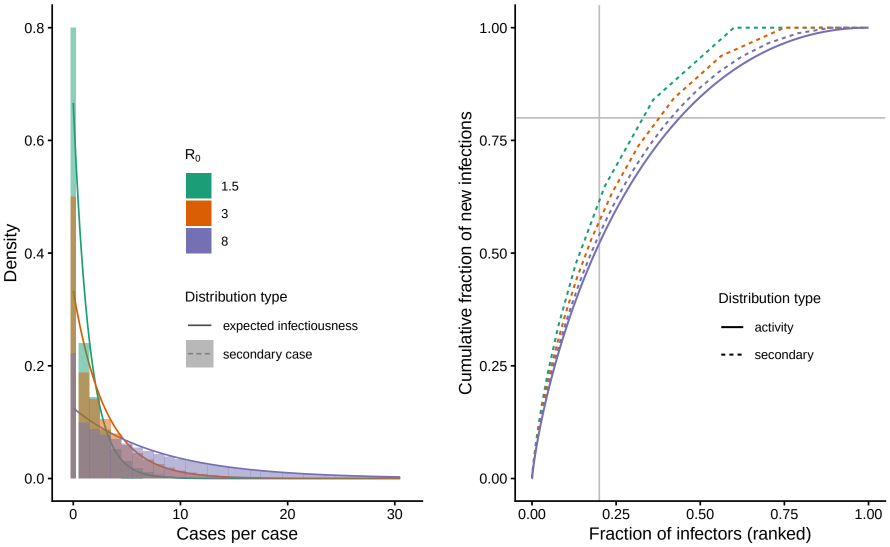

Figure 1: **Heterogeneity emerges even from a simple, linearized compartmental model** due to implicit variation in recovery times among infectors. (left) Activity distributions (density curves) and secondary case distributions (density histograms) for the outset of an SIR epidemic. Because the first bin (at zero) sits at the boundary of support for each distribution, we have plotted this bin as double the density and half the width; this adjustment preserves area-to-area correspondence with the PDF, while facilitating visual comparison of the heights of the density and mass functions. (right) Inequality curves for *activity* distributions from SIR models with differing $\mathcal{R}_0$ are identical (and indistinguishable due to overplotting); inequality in the *case* distribution decreases with R0 towards the theoretical limit of the activity distribution.

Figure 1 shows patterns of emergent heterogeneity in the early stage of an outbreak.

But despite differences in a non-dynamic world, we find invariance in case-per-case when looking across the entire epidemic. The top panel of Figure 2 shows realized distributions of "offspring cases" caused by individual infectors across a simple, stochastic SIR epidemic. The distributions remain indistinguishable across a wide range of the key parameter $\mathcal{R}_0$. This seems surprising. The resolution is that larger epidemics with larger $\mathcal{R}_0$ have larger between-cohort variation, as expected, but that is balanced by smaller within-cohort variation (Figure 2, lower panel).

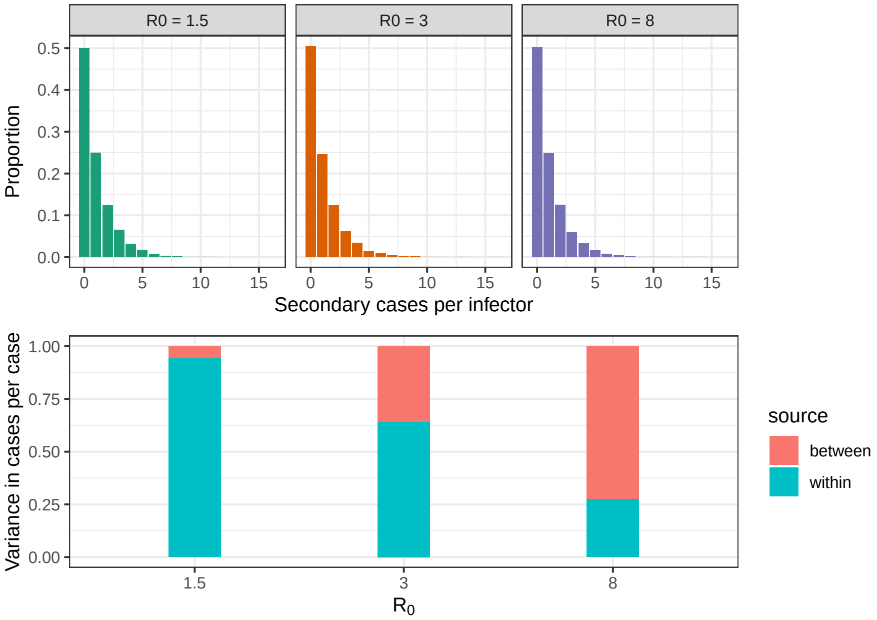

Figure 2: **Identical distributions of cases per case for epidemics of different strength conditioned over the entire outbreak cycle.** Epidemic outbreaks have been simulated for a population of 10000 with different $\mathcal{R}_0$, and data on the case per case have been collected at the end of the outbreak (top panels). Different compositions of variance in cases per case for epidemics of different strengths (lower panel). As $\mathcal{R}_0$ grows, between-cohort variation increases while within-cohort variation decreases, so that the total variance in cases per case remains constant.

We are also interested in what emergent distributions will look like to people studying outbreaks in real time. We are interested, at least to some extent, both in how cohorts change through time, and in what the outbreak will "look like" if we observe from a particular time. Figure 3 is one example; we are working on others.

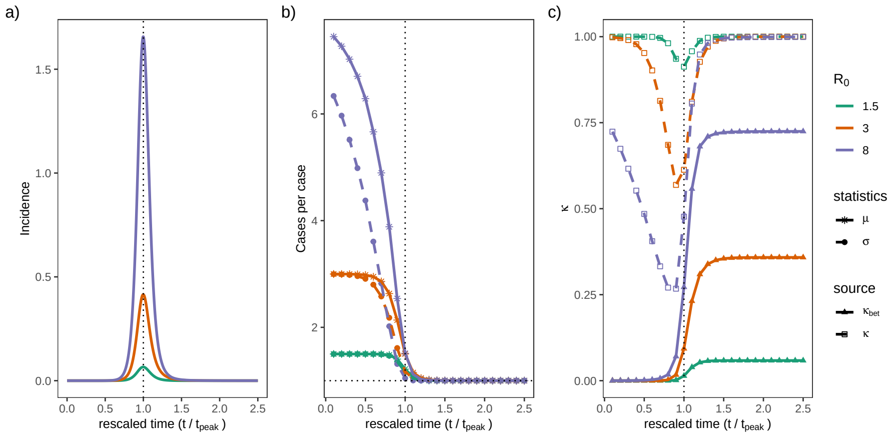

Figure 3: **The expected number of secondary cases generated by each case evolves over the course of the outbreak.** Early in the outbreak, the susceptible pool depletes slowly, and variation in cases per case comes mainly from variation in the recovery times, so the between-cohort component of variation is negligible (panel c). As the outbreak approaches its peak, depletion of the susceptible pool accelerates, making the length of the infectious period less important as a source of variation. Well after the peak, once the susceptible pool has become nearly constant again, variation in the infectious period once more drives the variation within cohorts. However, the share of within-cohort variation in the total variation is smaller in strong epidemics. The horizontal axis represents the *rescaled* time relative to the outbreak peak time. By "rescaled," we mean that time has first been scaled by the mean infectious period and is then measured relative to the outbreak peak. Panel a shows the evolution of the incidence over the course of the outbreak. Panel b shows the mean and standard deviation of the number of cases per case . Panel c shows the squared coefficient of variation, decomposed into within-cohort and between-cohort components. For panels b and c, the y-axis value at each time point was computed using only the cohorts that had been infected up to that time.

2025 Dec 09 (Tue) suggestion for figures. We want:

- a naive truncated figure that assigns to each cohort the number of actual cases up until a particular time.**[AA:** *Does Figure 4 do what we are looking for?***]**
- an idealized truncated figure that gets each cohort right (this is the current Figure 3), the idea is that it can also represent an idealized version of nowcast perceptions.
- A cohort-description figure but without cumulating for the supp. This one does not need to bother with between-cohort statistics. That is going to be time-scaled version of an older figure. **[AA:** *Does Figure S1 do what we are looking for?***]**

It's also possible to imagine realistic approaches between these two extremes, but let's put that off for later. There are methods (including by Dushoff and Park) for thinking about this at the cohort level, but not with a focus on individual variation. Maybe this is just for discussion. OR maybe we should also look at plots where we go up until a particular time and only count recovered infectors **[AA:** *Does Figure 5 do what we are looking for?***]**.

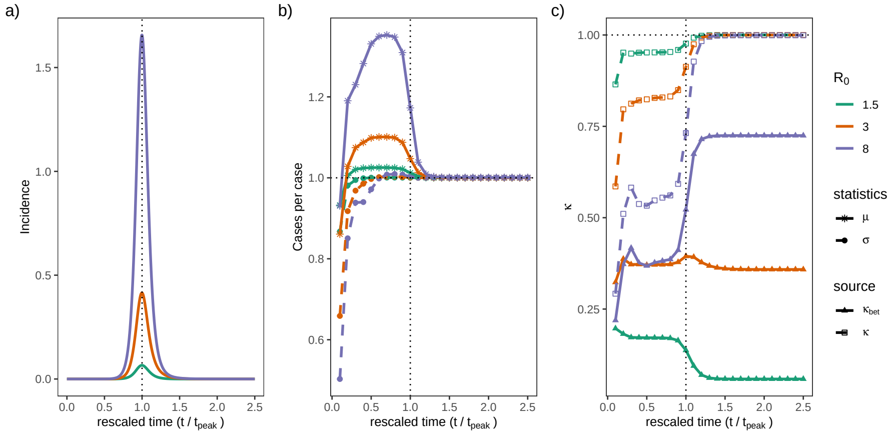

Figure 4: **The expected number of secondary cases generated by each case evolves over the course of the outbreak.** Panels (a-c) correspond to incidence, the mean and standard deviation of the number of cases per case, and its squared coefficient of variation, respectively, as in Figure 3. For panels b and c, the simulation was stopped at each time point to compute the y-axis value. The y-axis value at each time point was computed by assigning to each cohort the realized cases up until that time.

[TG: *Can we make a note about for epidemics with large R0, if you don't start tracking cases right from the beginning, you'll already underestimate cases/case* **]** **[JD:** *Yes, this should go into the paper.***]**

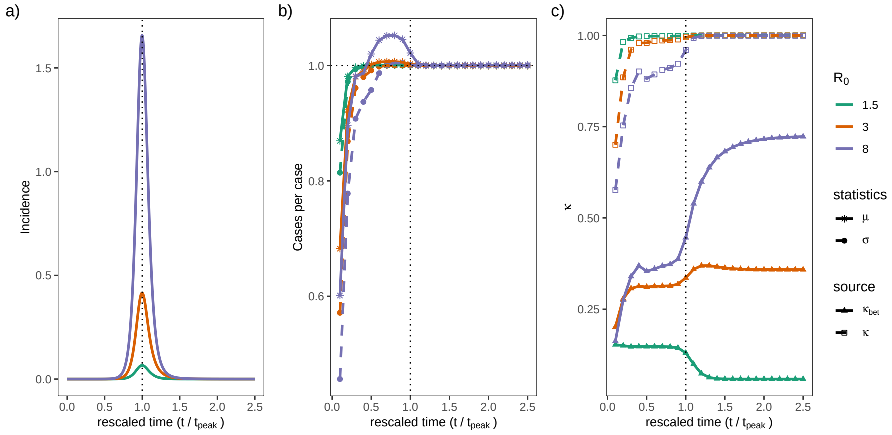

Figure 5: **The expected number of secondary cases generated by each case evolves over the course of the outbreak.** Panels (a-c) correspond to incidence, the mean and standard deviation of the number of cases per case, and its squared coefficient of variation, respectively, as in Figure 3. For panels b and c, the simulation was stopped at each time point to compute the y-axis value. The y-axis value at each time point was computed by taking into account the recovered individuals by that time. More specifically, the contribution of each cohort to the overall mean and variance was weighted by the fraction of recovered individuals in the cohort.

4 Box (or appendix?) Tapan's proof?

5 Discussion

6 Materials and Methods

The values of secondary cases in the left panel of Figure 1, were generated by computing the geometric probability density function with mean $\mathcal{R}_0$ at the points $0, 1, \ldots, 30$. As for the expected infectiousness, we evaluated the exponential probability density function with mean $\mathcal{R}_0$ at 300 equally distanced point in the interval $[0, 30]$.

In the right panel of Figure 1, we used the Lloyd-Smith's approach to compute the inequality in the activity distribution. More specifically, we used the relation $F_{\text{trans}}(x) = \frac{1}{\mathcal{R}_0^2} \int_0^x ve^{-\frac{1}{\mathcal{R}_0}} dv$ which is the fraction of cases due to those caused up to x secondary cases. The fraction of cases due to those caused more than x cases would be $1 - F_{\text{trans}}(x)$, which is equal to $(1 + \frac{x}{\mathcal{R}_0})e^{-\frac{x}{\mathcal{R}_0}}$. The population fraction of the individuals infected more than x cases is $e^{-\frac{v}{\mathcal{R}_0}}$. We used a similar approach to compute the inequality in the secondary case distribution: We used the relation $F_{\text{trans}}(x) = \frac{1}{\mathcal{R}_0} \sum_{v=0}^{x} vG(v, \mathcal{R}_0)$ which is the fraction of cases due to those caused up to x secondary cases. Here $G(v, \mathcal{R}_0)$ is a geometric distribution with mean $\mathcal{R}_0$. The fraction of cases due to those caused more than x cases equals $1 - F_{\text{trans}}(x)$. The population fraction of these individuals would be $1 - P(x)$, where $P(x)$ is the cumulative distribution function of a geometric distribution with mean $\mathcal{R}_0$ evaluated at x.

To generate the top panel of Figure 2, we used the individual based simulation and computed the case per case reproductive number in a population of size 10^4. The simulation was initiated with one case.

We used a deterministic SIR model to compute the mean and variance of $\mathcal{R}_{\text{c}}$. First, we scaled time by the mean infectious period, so the resulting SIR model then depends on one parameter: $\mathcal{R}_0$. The SIR differential equations were numerically solved for the proportions of susceptible $x(t)$ and infectious $y(t)$ at each time point t. The time interval for integration was set to $[0, 100]$, well after the outbreak died out. We then used the **R** function `approxfun` to construct a function, $X(t)$, that linearly interpolates the time-series $(t, \mathcal{R}_0 x(t))$. We associated a cohort to each of the first 60% time points in the time-series t (We used a 60% cutoff to ensure cohorts had almost recovered by the end of the simulation period) and calculated the cohort-specific mean and variance. More specifically, for the cohort infected at time point τ, the mean, $\mu(\tau)$, and variance $\sigma^2(\tau)$ of $\mathcal{R}_{\mathrm{c}}$ read as:

$$\begin{aligned}
\mu(\tau) &= \int_{t>\tau} f(t-\tau) \int_{\tau}^{t} \mathcal{R}_0 x(s) ds\, dt\,, \\
\mathbb{E}[\mathcal{R}_{\mathrm{c}}^2(\tau)] &= \int_{t>\tau} f(t-\tau) \left(\int_{\tau}^{t} \mathcal{R}_0 x(s) ds \right)^2 dt\,, \\
\sigma^2(\tau) &= \mathbb{E}[\mathcal{R}_{\mathrm{c}}^2(\tau)] - \mu^2(\tau).
\end{aligned}$$

To calculate these quantities for the cohort infected at τ, the interpolating function $X(t)$ was integrated from τ to t, which is the case reproductive potential associated with the cohort fraction recovered at time point t. The mean $\mathcal{R}_{\mathrm{c}}$ for each cohort $\mu(\tau)$ was then the integral of the case reproductive potential weighted by the infectious period density function. We used the same approach to compute $E[\mathcal{R}_{\mathrm{c}}^2(\tau)]$ and the variance of $\mathcal{R}_{\mathrm{c}}$ for each cohort $\sigma^2(\tau)$. In Figure S1, the middle panel was generated using this approach.

We obtained the mean of $\mathcal{R}_{\mathrm{c}}$ by integrating the cohort-specific mean $\mu(\tau)$ against the incidence, $i(t) = \mathcal{R}_0 x(t) y(t)$, and normalizing the result by the final size, $\int i(t)dt$. We computed the between variance by integrating over the cohort's mean $\mu(\tau)$ weighted by the incidence. The result was then normalized by the final size. The with-in cohort variance was computed by taking integral over the cohorts' variance $\sigma^2(\tau)$ weighted by the incidence. The result was also normalized by the final size. The total variance was the sum of the between and with-in variances. We also calculated the total variance independently; we integrated $\mathbb{E}[\mathcal{R}_{\mathrm{c}}^2(\tau)]$ weighted by the incidence and divided it by the final size. The result minus the squared mean of $\mathcal{R}_{\mathrm{c}}$ yielded the total variance. Both approaches produced the same value for the total variance.

All integrations were done using `lsoda` method with the **R** package `deSolve`. All simulations were carried out using **R** `4.5.2`. Code for all numerical simulations is housed at: `https://github.com/dushoff/kappaCode`.

We solved all integrals across a range of values for $\mathcal{R}_0$, using the starting values $y_0 = 10^{-9}; x_0 = 1 - 10^{-9}$ to represent the limiting case in which there are no exogenous cases. In building these simulations, we used a range of time step sizes, noting convergence towards known and conjectured values (e.g., epidemic final size, mean case reproduction number, variance in case reproduction number) as resolution increased.

7 Supplementary Materials

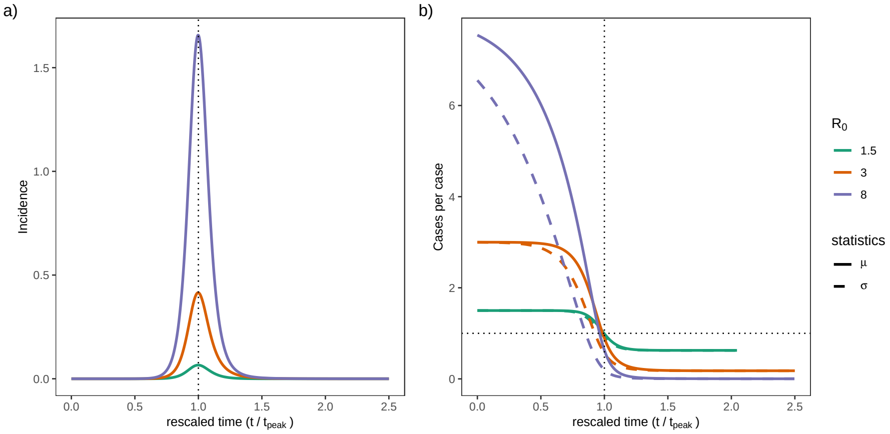

Figure S1: **Between-cohort variation in $\mathcal{R}_c$ increases as $\mathcal{R}_0$ increases.** Panel a depicts the cohort size against the cohort time of infection in units of mean infectious period divided by the incidence peak time; the cohort size is the same as the incidence. In Panel b, the mean and standard deviation of $\mathcal{R}_c$ associated with each cohort are plotted. In an epidemic with $\mathcal{R}_0$ close to 1, there is little susceptible depletion well after the outbreak onset, and the variation in infectious period mainly drives the variation in $\mathcal{R}_c$. Panel c plots the squared coefficient of variation, κ. Around the time of peak incidence, κ reaches its minimum, then rebounds towards one–taking much longer (in rescaled time) in stronger epidemics (larger $\mathcal{R}_0$).