1 Introduction

Box 1. κ tutorial.

For a continuous distribution with mean μ and variance σ^2, we define (generalized) *dispersion parameter*, κ, as $\frac{\sigma^2}{\mu^2}$. We define the dispersion parameter for a discrete distribution with mean μ and variance σ^2 as $\kappa_{\text{discrete}} = \frac{\sigma^2-\mu}{\mu^2}$.

These statistics have desirable properties for a dispersion measure: they *increase* with dispersion around the mean and equal zero for a Dirac delta function **[AA:** *How is* κ_{discrete} *zero for a* discrete *Dirac delta distribution?***]**.

The dispersion parameter is invariant under Poissonification: If a gamma distribution is parameterized by its mean μ and dispersion κ, then the negative binomial distribution arising from the Poisson mixture with that gamma is also parameterized by μ and $\kappa_{\text{discrete}} = \kappa$.

More generally, the connection between κ and κ_{discrete} holds for the Poisson mixture with any underlying distribution. Thus, we can use κ statistics to compare models with different mean-variance structures.

In our setting, the continuous distribution corresponds to the expected transmission rate, and the discrete distribution arises from the stochastic realization of that rate, i.e., by compounding a Poisson distribution with a rate parameter derived from the continuous distribution. Hence, κ can be used to characterize the dispersion of secondary case distributions.

2 Results

Demographic stochasticity can generate "emergent" heterogeneity even in the absence of explicit differences between individual-based rates (Figure 1). 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 shows patterns of emergent heterogeneity in the number of secondary cases during the early stage of an outbreak,i.e., in a fully susceptible population. Both activity distribution and the secondary case distribution depend on $\mathcal{R}_0$. How would the secondary distribution look if we examined all infectors over the entire outbreak? To address this, we simulated epidemic outbreaks in a population of 10000 and, for each infected individual, recorded the number of secondary cases they generated throughout the outbreak. The top panel of Figure 2 shows realized distributions of "offspring cases" caused by individual infectors across a simple, stochastic SIR epidemic. Despite a non-dynamic world, the distributions remain indistinguishable across a wide range of the key parameter $\mathcal{R}_0$. This seems surprising.

To resolve this apparent discrepancy, we examined how heterogeneity differs when analyzing epidemiologically relevant cohorts, which are groups of individ-

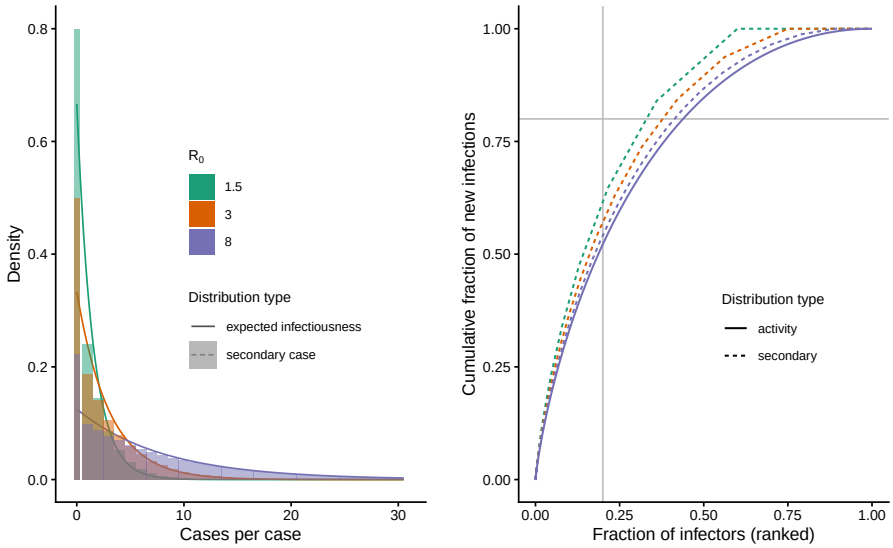

Figure 1: **Heterogeneity emerges even from a simple, linearized compartmental model** due to implicit variation in recovery times among infectors. (left) Activity distributions (density curves) and secondary case distributions (density histograms) for the outset of an SIR epidemic. Because the first 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 $\mathcal{R}_0$ towards the theoretical limit of the activity distribution.

uals infected at the same time (Figure S1). By doing so and splitting the total case-per-case variation into between- and within-cohort components, we found that 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).

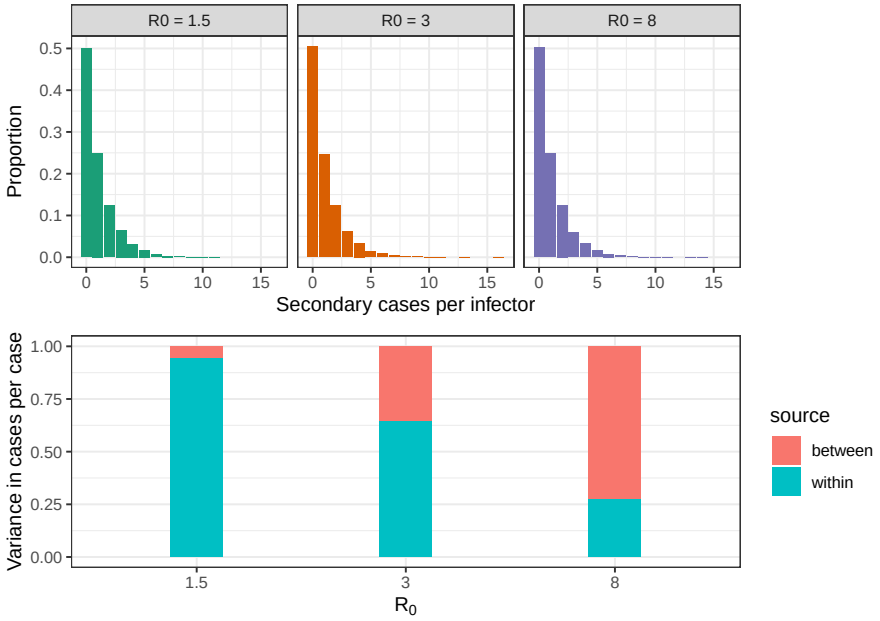

Figure 2: **Nearly identical empirical distributions of case-per-case for epidemics of different strengths conditioned over the entire outbreak cycle.** Epidemic outbreaks have been simulated for a population of 10000 with different $\mathcal{R}_0$ and the number of secondary cases per infector has been recorded at the time of outbreak extinction (top panels). Identical variance in case-per-case for epidemics of different strengths modeled using simple compartmental models (lower panel). As $\mathcal{R}_0$ increases, between-cohort variance rises, and within-cohort variance falls, maintaining constant total variance of 1.

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. To do so, we simulated a deterministic SIR model and stopped it at various cutoff points. At each point, we recorded the number of cases generated by each infected cohort up to that time. In the early stage of the outbreak, observed heterogeneity is lower compared to both a fully susceptible population (linearized model) and the entire outbreak (Figure 4). By the end of the outbreak, the mean and standard deviation of case-per-case

for different values of $\mathcal{R}_0$ become indistinguishable.

We also examined an idealized scenario in which we computed the case-per-case distribution among cohorts who got infected up to a given cutoff point. In epidemics with lower $\mathcal{R}_0$, the case-per-case distribution across early-infected cohorts is the same as that in a constant-susceptible pool (Figure 3). Incorporating cohorts who later, but prior to the outbreak peak, became infected reduced variability. Further incorporating those cohorts that became infected after the peak increased the variability. By the end of the outbreak, all epidemics with different $\mathcal{R}_0$ agree on the case-per-case distribution.

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.

[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.*]

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

4 Discussion

5 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 $[0, 30]$.

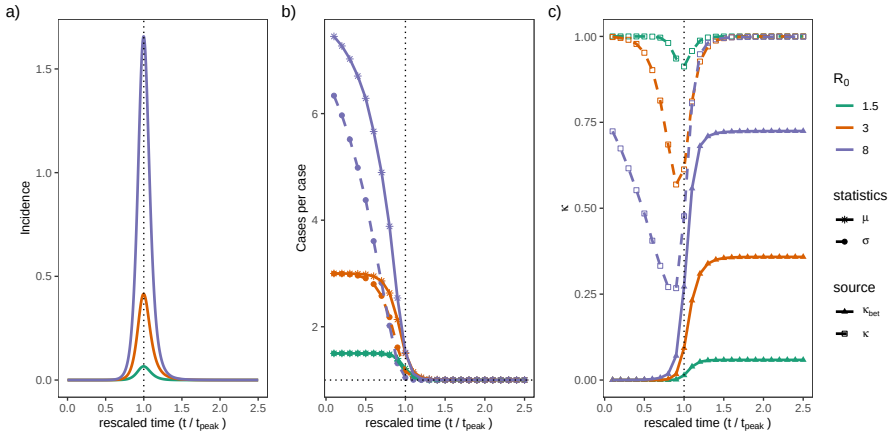

Figure 3: **Nowcasting the case-per-case distribution evolves as the outbreak unfolds.** Panel a: Incidence represents the size of each cohort infected at each rescaled time point. By "rescaled," we mean that time has first been scaled by the mean infectious period and is then measured relative to the outbreak peak. Panel b: At each time, the mean and standard deviation of cases caused by each case are calculated using only cohorts infected up to that time. The mean cases per case is larger in early-infected cohorts and depends on the key parameter $\mathcal{R}_0$. As time goes by, later-infected cohorts have larger weights, as measured by incidence. These cohorts experience susceptible pool depletion, leading to fewer secondary cases and a lower mean case-per-case number. After the peak, the susceptible population size stabilizes, and the mean and variance approach 1. Panel c: The between-cohort component of the squared coefficient of variation κ_{bet} is negligible among the early-infected cohorts, as they experience a similar size of susceptibles. As the outbreak unfolds, the decline in the susceptible population drives between-cohort variation, which is more pronounced in epidemics with larger $\mathcal{R}_0$. The squared coefficient of variation κ reaches its minimum around the outbreak peak. The largest cohort, measured by incidence, experiences a sharp depletion of susceptibles, and the impact of variation in recovery time fades—declining the variance and, in turn, κ. After the peak, κ rebounds to one. In panel c, the squared coefficient of variation is shown, split into within-group and between-group 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.

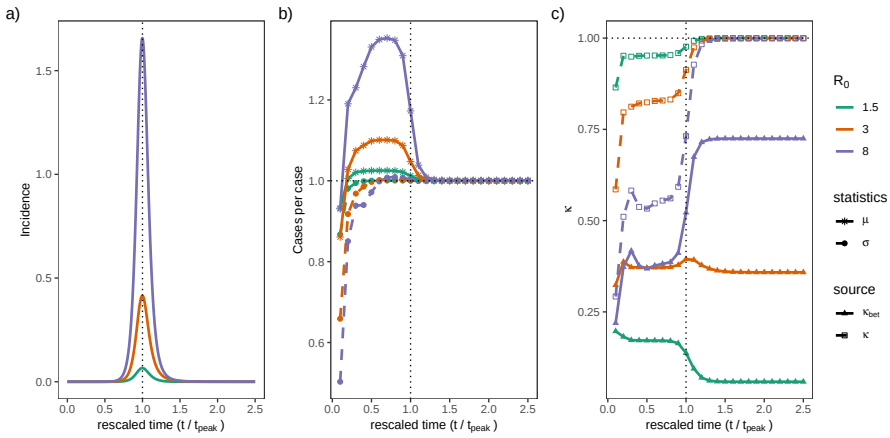

Figure 4: **Distribution of case-per-case evolves over the course of outbreak.** Panel a: The size of the cohort infected at each rescaled time point is measured by incidence at that point. Panel b: The mean and standard deviation of case-per-case at each time are computed using the realized cases up to that point. Early in the outbreak, the variability in case-per-case is yet to be fully revealed, as those still infectious generated the same number of secondary cases as those who had just recovered. As the outbreak approaches its peak, the contribution of the early infected but not yet recovered population is realized, increasing the standard deviation. After the peak, both the mean and the standard deviation approach 1 rapidly, regardless of $\mathcal{R}_0$. Panel c: The squared coefficient of variation of case-per-case κ climbs over the course of the outbreak. In epidemics with large $\mathcal{R}_0$, as opposed to those with $\mathcal{R}_0$ close to one, the between-cohort component jumps after the outbreak peak as the susceptible pool depletes. 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 number of realized cases up to that time. In all panels, time units measured in the mean infectious period are again scaled relative to the peak time.

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 $\mathcal{R}_{\text{c}}$ read as:

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

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 $\mathcal{R}_{\text{c}}$ 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 $E[\mathcal{R}_{\text{c}}^2(\tau)]$ and the variance of $\mathcal{R}_{\text{c}}$ for each cohort $\sigma^2(\tau)$. In Figure S1, the middle panel was generated using this approach.

We obtained the mean of $\mathcal{R}_{\text{c}}$ 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 $\mathbb{E}[\mathcal{R}_c^2(\tau)]$ weighted by the incidence and divided it by the final size. The result minus the squared mean of $\mathcal{R}_c$ 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.

6 Supplementary Materials

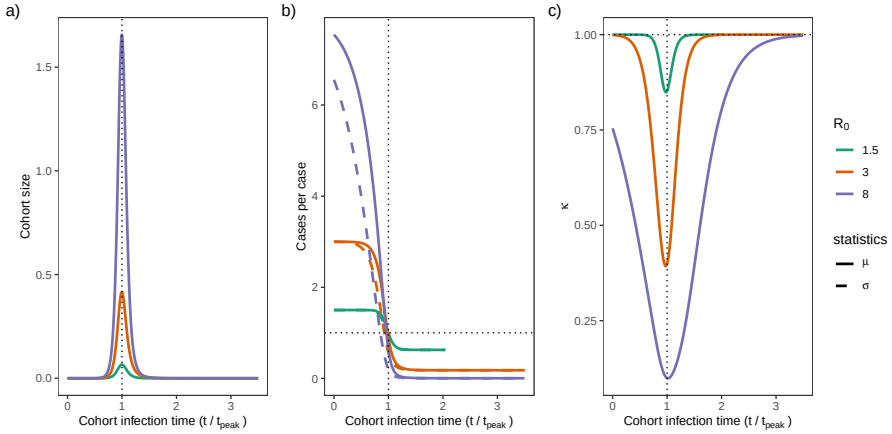

Figure S1: **Between-cohort variation in $\mathcal{R}_c$ increases as $\mathcal{R}_0$ increases.** Panel a depicts cohort size as a function of the time of infection, in units of the mean infectious period divided by the incidence peak time; the cohort size equals the incidence. In Panel b, the mean and standard deviation of $\mathcal{R}_c$ associated with each cohort are plotted. Early in the outbreak, the susceptible pool shrinks slowly, and differences in recovery time mainly cause variations in case numbers. As a result, the average for cases is about the same as in a simple linearized model. Near the outbreak's peak, faster susceptible loss partly compensates for the inequality stemming from the infectious period, and, in turn, the squared coefficient of variation κ decreases (Panel c). After the peak, once the susceptible pool has become nearly constant again, variation in the infectious period once more drives variation within cohorts; κ approaches 1. In stronger outbreaks (higher $\mathcal{R}_0$), it takes longer (in rescaled time units) for the susceptible population to stabilize, explaining the slower move of κ toward one.