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 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).

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.

Observing from a particular time can be done in two ways: either naively, or by trying to correct for the truncation of observations. These can be simulated, respectively, by either simply stopping the simulation at a certain time **[AA:** *Does Figure 5 do what we are looking for?***]** (or reporting what would be seen if we did), or in an idealized world, by looking at all the cohorts infected up until a given time **[AA:** *Does Figure 3 do what we are looking for?***]**. It's worth looking at some pictures of both of these views and seeing what we think. It may also be worth looking at statistics for individual cohorts (I guess this is a bit boring, because we only have within-cohort variation in that case, but we should do it and put in the supp).

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 4 do what we are looking for?***]**.

[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*

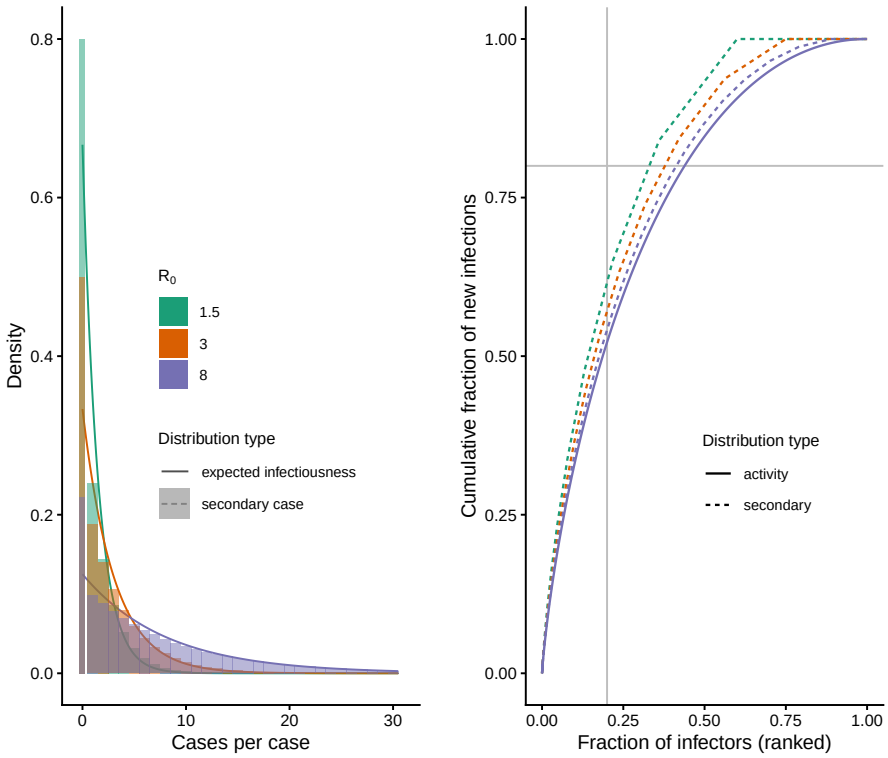

Figure 1: **Heterogeneity emerges even from a simple, linearized compartmental model** due to implicit variation in recovery times among infectors. (left) Activity distributions (density curves) and secondary case distributions (density histograms) for the outset of an SIR epidemic. Because the first 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.

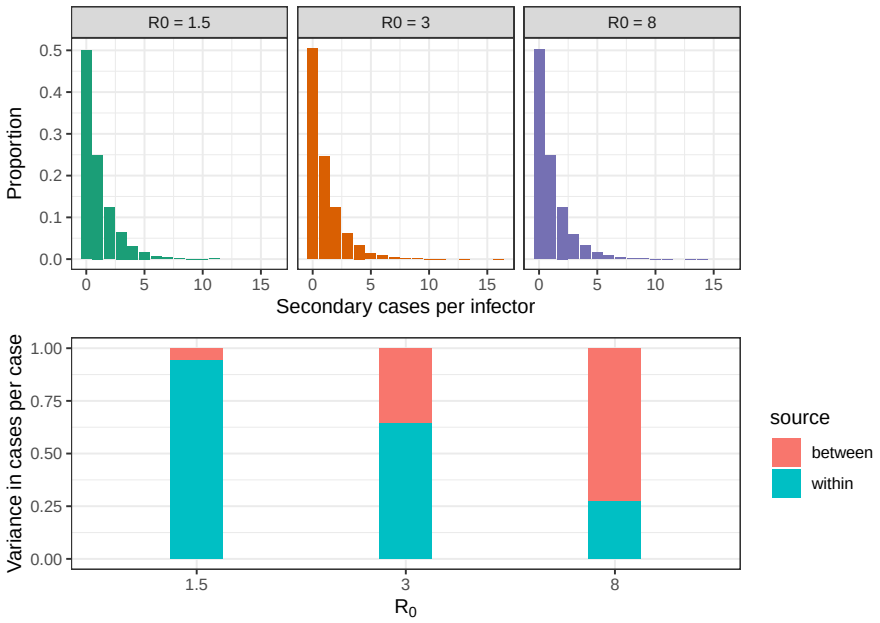

Figure 2: **Identical distributions of cases per case for epidemics of different strength conditioned over the entire outbreak cycle.** Epidemic outbreaks have been simulated for a population of 10000 with different $\mathcal{R}_0$, and data on the case per case have been collected at the end of the 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.

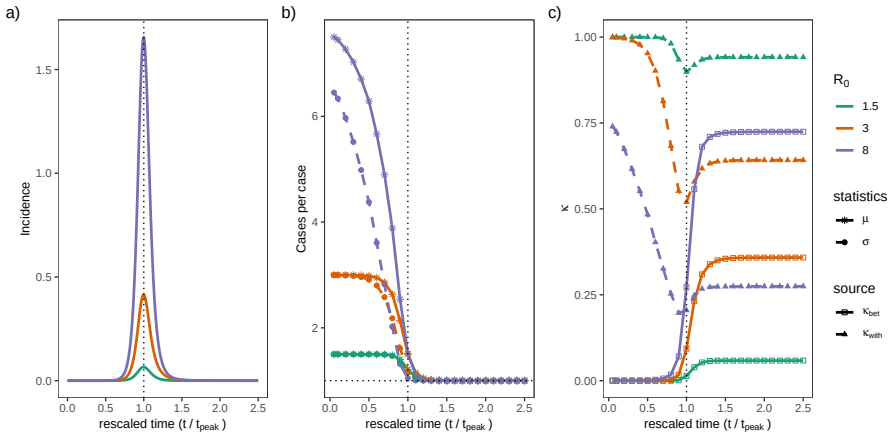

Figure 3: **The expected number of secondary cases generated by each case evolves over the course of the outbreak.** Early in the outbreak, the susceptible pool depletes slowly, and variation in cases per case comes mainly from variation in the recovery times, so the between-cohort component of variation is negligible. 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. 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 is computed using only the cohorts that have been infected up to that time.

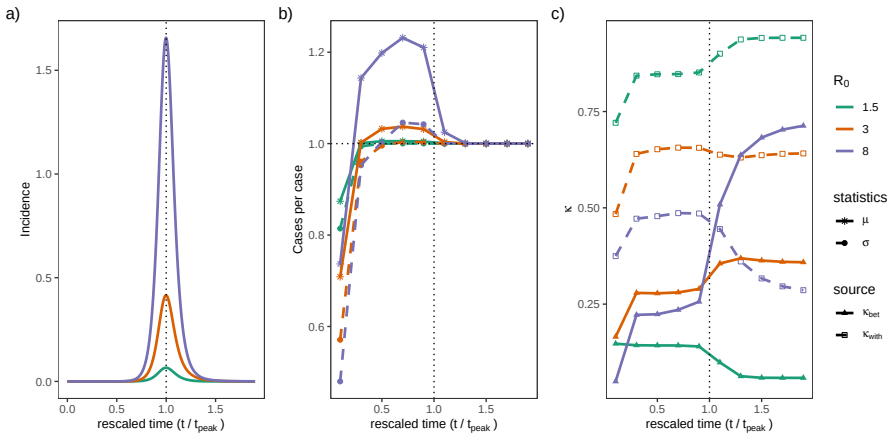

Figure 4: **The expected number of secondary cases generated by each case evolves over the course of the outbreak.** For panels b and c, the y-axis value at each time point is computed by counting recovered infectors.

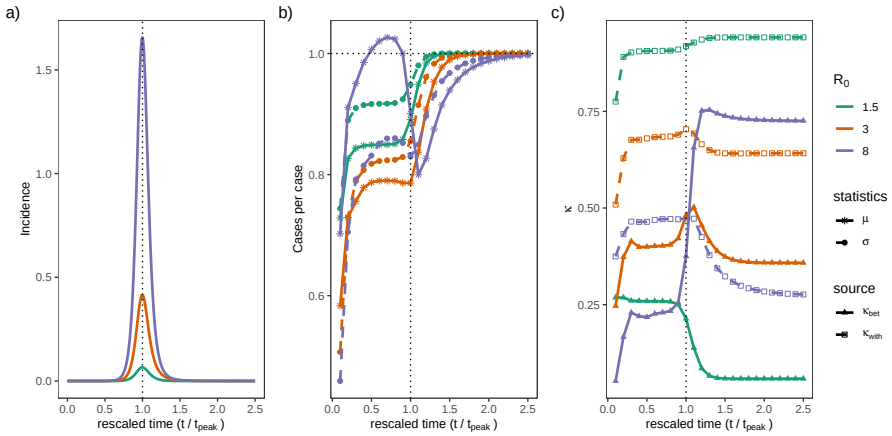

Figure 5: **The expected number of secondary cases generated by each case evolves over the course of the outbreak.** For panels b and c, the simulation was stopped at each time point to compute the y-axis value. At each time, we calculated, for each cohort, the mean and variance of the number of cases per case among recovered individuals by that time, and used these values for the whole cohort. Hence, unlike in Figure 4, where each cohort's contribution to the overall mean and variance of the number of cases per case was weighted by its recovered fraction, here cohort s are represented by the statistics of their recovered subsets at that time point.

] [**JD:** *Yes, this should go into the paper.*]

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

5 Discussion