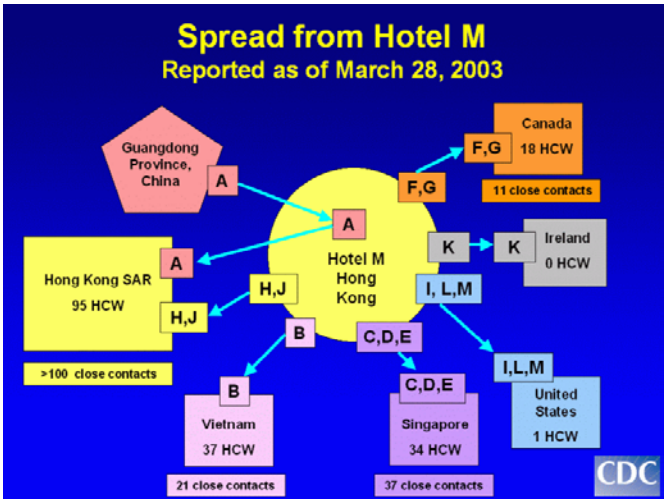
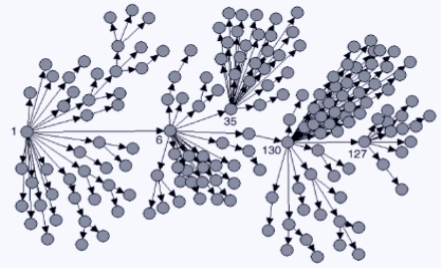
The individual reproductive number in the study of epidemics

Ekkehard Kopp, SACEMA, University of Stellenbosch, South Africa



This paper, based on an invited talk given at the 2007 SAMSA meeting in Windhoek, Namibia, is purely expository, and seeks to highlight novel epidemiological techniques introduced in [2] and commented on extensively in [1]. The motivation for this research was better to understand the nature of epidemics of infectious ('casual contact') diseases which exhibit a high degree of 'superspreading', such as the SARS epidemic which occurred in 2002/3. The unprecedented global effort to contain SARS generated unique datasets, characterized by intensive contact tracing, such as the one below, taken from [2]



The analysis of such data sets proved challenging for standard statistical modelling techniques. The present paper seeks to bring the extension of such techniques initiated in [2] to a mathematically oriented audience, and, in particular, presents ideas from the web-based Supplementary Material accompanying [2]. No claims of originality are made here: all the information can be found in the above sources. The pictures and graphs reproduced here are taken from slides provided by Jamie Lloyd-Smith (private communication).

Superspreading events for SARS highlighted the critical importance of individual variation for disease emergence and thus pointed to apparent shortcomings in existing epidemiological theory.

The **basic reproductive number, R_0** , i.e. the expected number of secondary cases caused by a typical infectious individual in a completely susceptible population, has long served as a key indicator in the study of epidemics. In particular:

$R_0<1$ → the disease will die out

$R_0>1$ → the disease can invade the population

However, it is fundamentally an average measure, neglecting heterogeneity.

The elementary theory of **branching processes** provides simple tools for the study of infectious diseases. Starting (at time 0) with a single infectious individual we have the following key ideas:

- The offspring distribution: p_k =Pr(Z=k), k=0,1,2,… Z is the number of 'offspring' of (infected by) an individual in a (large and) wholly susceptible population. The main tool is the probability generating function (pgf) f: its slope f'(1) gives the mean R_0 of Z, the y-intercept f(0) gives the proportion p_0 causing no secondary infections.

- The size of the n^{th} generation Z_n has pgf F_n given by the n^{th} iterate of f, so that the probability of extinction in this generation is $f_n(0)$. In the limit the extinction probability is therefore q, the unique solution of the functional equation f(q)=q.

- $R_0>1$ provides a unique q<1 (else q=1 is the only solution).

- Let $Y= Z_0 + Z_1 + Z_2 +\ldots$ denote the total number infected during the epidemic, then its generating function F will satisfy: F(s)=sf(F(s)).

- Define a minor outbreak as one which goes extinct. Then its expected size is F'(1), which can be found numerically for a given f.

- Demographic stochasticity in transmission (contacts between infectives and susceptibles leading to secondary infection) is modelled by a **Poisson process**.

- Thus in a homogenous population f(s)=exp(-R_0(1-s)). If R_0 is replaced by a ***random variable*** ν with density f_ν the offspring distribution becomes a **Poisson mixture**.

Using a random variable ν to replace R_0 reflects the fact that transmission of infection through casual contact is affected by a large number of factors which are not easily parametrised and quantified. Net infectiousness of each host is determined by combination of host, pathogen, and environmental factors, for example:

- Contact rate
- Probability of transmission per contact (arguably related to pathogen load?)
- Host behaviour, hygiene
- Infectious period
- Virulence/transmissibility of strain
- Local population density and susceptibility
- Control measures, time to isolation, medical treatment
- Circumstance and fluke events.

Constructing the offspring distribution

We consider three candidate models for the offspring distribution which can be tested against the above-mentioned data sets: in each case let ν_i be the expected number of secondary cases caused by individual *i.* Demographic stochasticity in the transmission process is modelled by Z ~

Our three cases are characterized by the probability distribution assumed for ν_i

1. Assume that $\nu_i=R_0$ is constant – this is the classical assumption used in the literature. The offspring distribution is then $Z \sim \text{Poisson}(R_0)$

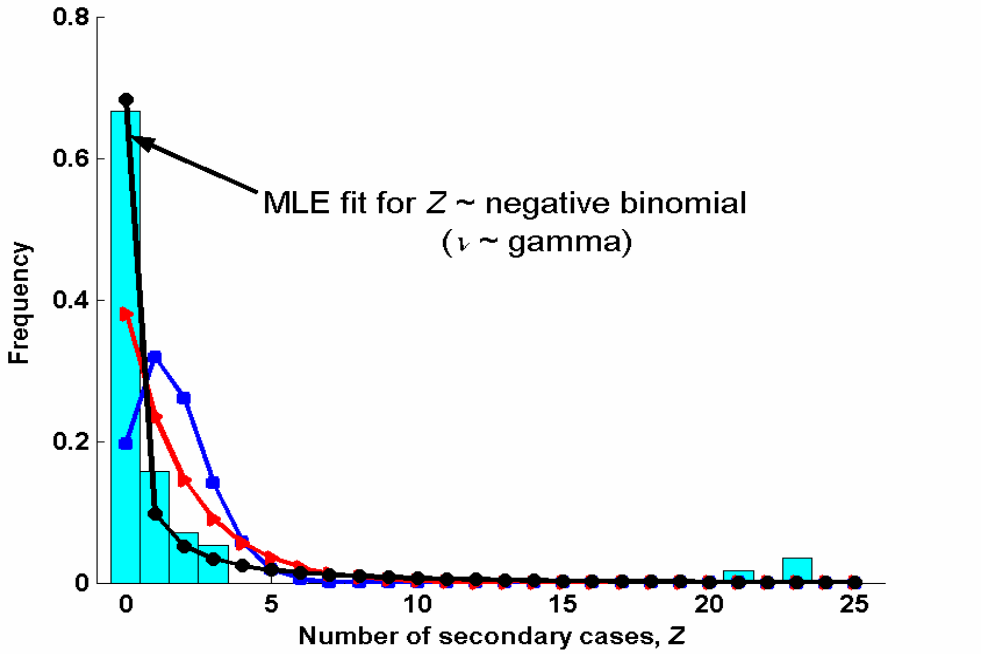
2. Assume a constant recovery rate and homogeneous infectiousness. Now ν is exponentially distributed, with mean R_0 so that $Z \sim \text{Geom}(R_0)$ and the Pgf of the offspring distribution is $f(s)=(1+R_0(1-s))^{-1}$

3. Allowing a more flexible degree of variability in individual infectiousness (at the expense of an additional parameter) let ν be Gamma-distributed, with mean R_0 and shape parameter k. Hence $Z \sim \text{NegB}(R_0, k)$ has a negative binomial distribution. The Pgf of the offspring distribution is then $f(s)=(1+(R_0/k)(1-s))^{-k}$

Note that we use the *mean* instead of the *scale parameter* to compare the three cases: *Po*(R_0)*; Geom*(R_0)*; NegB*(R_0, k) refer to the distributions with conventional notation *Po*($1/R_0$)*; Geom*(*p*) with $p = (1 + R_0)^{-1}$ and *NegB*(*p;k*) with $p = (1+ R_0/k)^{-1}$ respectively.

The graph below illustrates MLE fits to the data for the Singapore SARS outbreak in 2003 against these three candidate models. The red and blue graphs give the fit for models 1 and 2 respectively. It is clear that, with an appropriate value for k (which turns out to be k=0.16 in this case) by far the best fit is achieved with model 3.

Singapore SARS outbreak, 2003



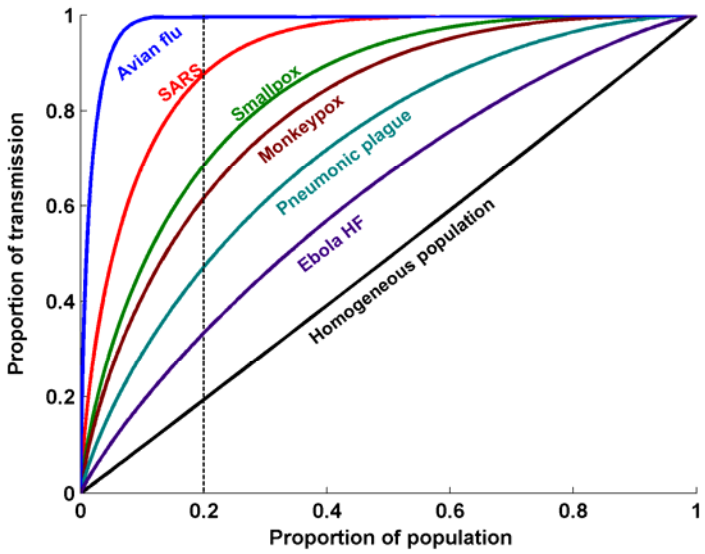
Similar conclusions can be drawn using the Akaike test (see [2] for details) to compare these three models:

v distribution	Z distribution	DAICc	Akaike weight
Constant	Poisson	250.4	< 0.0001
Exponential	Geometric	41.2	< 0.0001
Gamma	Negative binomial	0	>0.9999

Model selection strongly favours the NB model with R_0 =1.63 and k=0.16. Graphs produced in [2] illustrate similar calculations for a range of other diseases of casual contact, with somewhat mixed results. Inevitably, with quite small data sets, confidence intervals tend to be large. While few general conclusions can be drawn from this exercise, it does indicate clearly that variability in infectiousness is present in all these cases, and that its modelling requires care.

Where inhomogeneity in infectiousness has been addressed in the literature (especially for vector-borne diseases), a popular assumption has been an '80:20 rule': that 80% of infections are caused by the most infectious 20% of the infected population. In [2], this rule was tested for a variety of diseases of casual contact, with the following results, which show that the 'rule' should not be applied indiscriminately:

Revisiting the 20/80 rule (see [2] for details of the datasets employed)



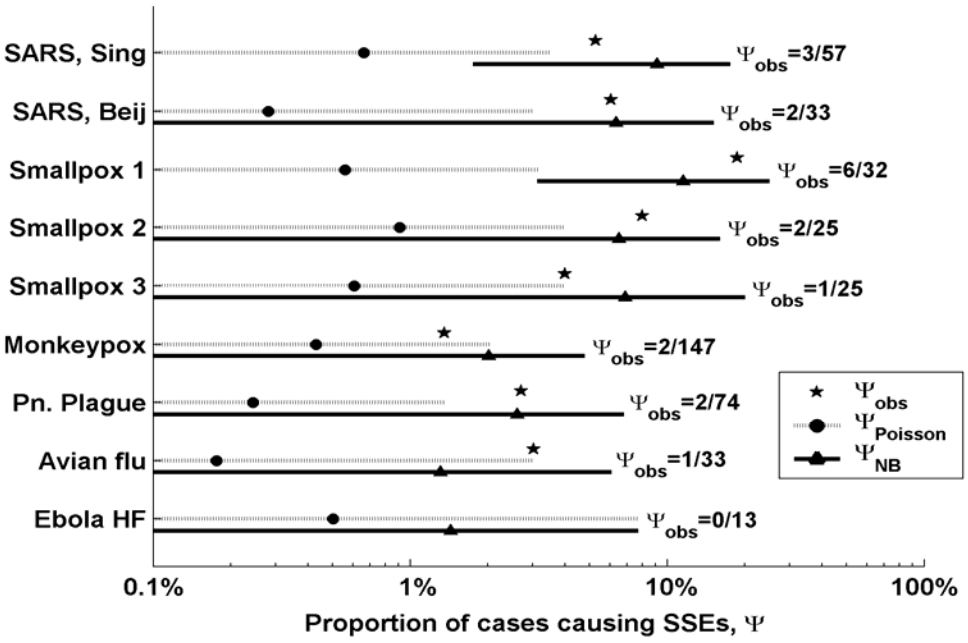
Clearly, more and larger datasets are needed to refine this work. In particular, for diseases such as SARS, where the proportion of infected individuals causing no secondary infections is high, it would be beneficial to record this number, as estimates or the mean and the proportion of zero transmitters would allow modellers to provide effective tests of variability in infectiousness for such sar

What is a superspreading event?

A simple definition for superspreading events is provided in [2]. It stands in some contrast to earlier attempts, which either sought to ignore them as outliers, or provided arbitrary numbers for specific diseases.

In order to define a superspreading event for a given outbreak, estimate the effective value of R_0 from the data in question, then generate $Z \sim \text{Poisson}(R_0)$ to provide the offspring distribution in the absence of variability in infectiousness, and finally designate an event as a superspreading event (SSE) if it lies in the top percentile, i.e the individual has infected more than $Z^{(99)}$ others, where $Z^{(99)}$ denotes the 99th percentile of the Poisson distribution.

This idea was tested against reported SSEs for a number of diseases where datasets are available, and showed a high degree of correspondence. The negative binomial (NB) model and the homogeneous Poisson model were then tested against observed data to establish their predicted proportions of SSEs. The superiority of the NB model was again evident, giving a close fit with reported events in most cases.

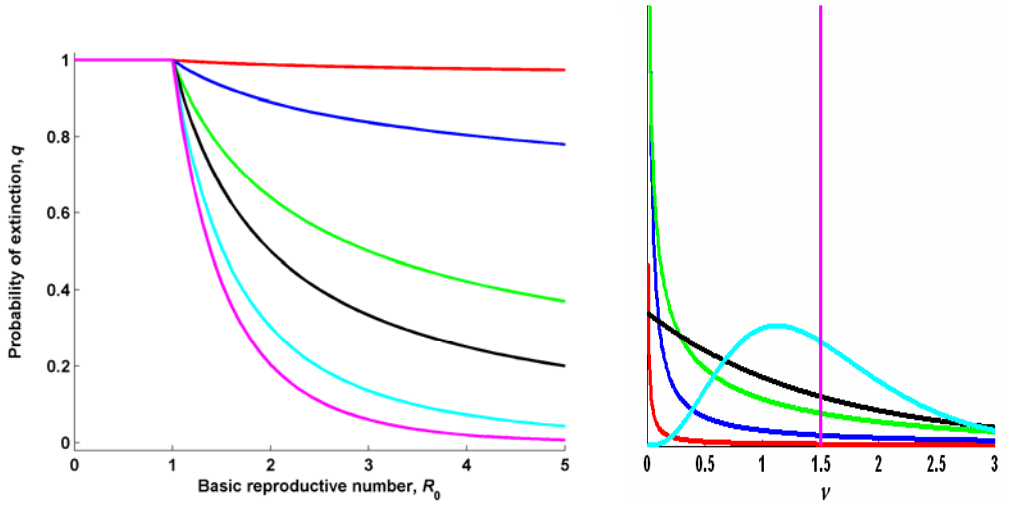


Impact upon Disease Control policies

Data from 10 diseases of casual contact have thus shown that individual variability in ν is a universal phenomenon. How does this variability affect:

- Probability of stochastic extinction
- Timing of extinction
- Size of minor outbreak
  (number of cases before extinction)
- Rate of growth if outbreak occurs

We can explore these questions for a branching process model with ν ~ gamma. First, the probability of disease extinction can be plotted against different values of the shape parameter k – smaller k implies greater variability:



In this graph, the different colours represent different values of k, ranging from 0.1 (red) to infinity (purple – constant infectiousness). It is evident that for small k, the disease will die out with high probability and will do so in a few generations. However, when it does take hold, it will spread explosively – again, see [2] for simulation graphs and details.

The control policies examined in [2] are homogeneous partial (HP) control (i.e. everyone's infectiousness is reduced by a fixed factor c , as, for example, through vaccination with a partially effective vaccine), and random absolute (RA) control (such as through isolation) where a proportion c of the population has its infectiousness reduced to 0. Using gamma-distributed infectiousness, this leads to the following:

For HP control, with all individuals' transmission reduced by a factor *c*, the offspring distribution is Z(HP,c)~NegB((1−*c*)R_0,*k*) and has pgf:
$f_{HP}(s)=[1+(1-c)(1-s)R_0/k)]^{-k}$

For RA control, with a random proportion *c* of individuals controlled absolutely, the pgf is:
$f_{RA}(s)=c+(1-c)[1+ (1-s)R_0 /k]^{-k}$

This follows because the control affects transmission only for the fraction $(1-p_0)$ of individuals whose natural Z value is greater than zero. A fraction *c* of these then have an RA-controlled Z-value of 0, and that of the remainder remains unaffected.

So the proportion of cases causing zero infections becomes $p_0+c(1-p_0)$, and the population mean becomes $(1-c)R_0$.

Thus the zero-class for the composite distribution is expanded, while for j>0 the density is simply reduced by a factor *c* from the original negative binomial. So offspring distribution with RA control has pgf
$c+(1-c)[1+(1-s)R_0/k]^{-k}$

So that its variance to mean ratio equals $1+(R_0/k)+cR_0$, which increases monotonically with c.

We can approximate this distribution by a negative binomial with mean $R=R_{RA}=(1-c)R_0$ and shape parameter $k=k_{RA}$ estimated as the solution of the equation $p_0+c(1-p_0)=(1+R/k)^{-k}$.

This leads to the following (see [2])

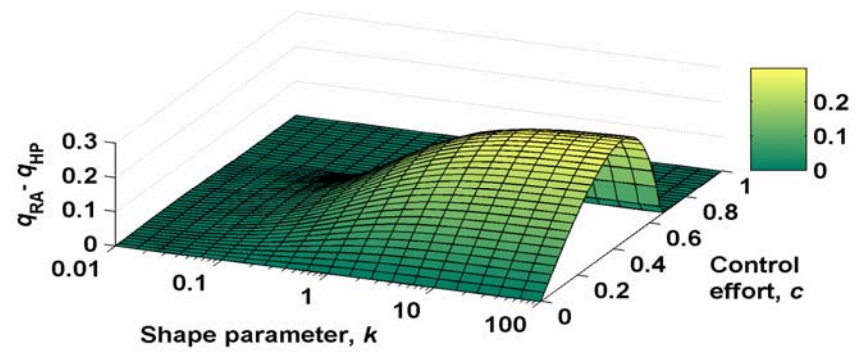
Proposition: For all c in $(0,1-1/R_0)$, and the probability of extinction is always greater under RA control than under HP control.

Proof: Define
$G(x)=[1+(xR_0)/k(1-s)]^{-k}$
where X is a Bernoulli random variable with probability $1-c$ of success.

Since G is a convex function, Jensen's inequality implies that
$f_{HP}(s)=G(E(X))<E(G(X))=f_{RA}(s)$ (*) for all c in (0,1) and s in [0,1).

Furthermore, for the n^{th} iterates of the pgf we have from (*) that
$f_{n,HP}(0)<f_{n,RA}(0)$ for all n,
so the probability of disease extinction by the n^{th} generation is always greater under RA control.
Thus if c is in $(0,1-1/R_0)$, the probability of ultimate extinction under RA control is greater than that under HP control, i.e. $q^{RA} > q^{HP}$.
If $c > 1-1/R_0$, then both R values are <1, so that $q^{RA} = q^{HP} = 1$. This completes the proof

One may also show that that for a given reduction in R_0, RA control is always more effective than HP. Thus it is better to completely control 50% of cases than to reduce transmission by 50% for all cases. This is depicted graphically, for different values of k, below:



To consider the efficacy of control policies targeting the more infectious individuals in a population, we consider a general branching process whose pgf is given in terms of the pdf $f_V(u)$ of the individual reproductive number for the outbreak in question.

For a control strategy C : $[0,\infty) \rightarrow [0,1]$ in which the probability of absolutely controlling an individual with individual reproductive number v is $C(v)$, the pgf of the branching process can be written down similarly for various choices of C.

It seems clear (and can be proved rigorously, see the Supplementary Material supplied with [2]) that for two control strategies with the same fraction of controlled individuals the one targeting more infectious individuals will be more effective. If both adjusted mean reproduction numbers are >1, the probability of extinction is greater with the more targeted control.

Conclusion

While the results in [2], as reported here, are restricted to simple models and do not address in detail a realistic set of control policies, the importance of ensuring that mathematical models are tested against available data sets in order to better inform the models, cannot be overemphasised. The results described here show that individual variability in infectiousness needs to be addressed in modelling the early stages of disease outbreaks, where effective diagnosis and control measures are not in place – such as was seen with SARS and may occur with species-crossing outbreaks.

References

1. Alison Galvani and Robert May: *Dimensions of Superspreading,* Nature 438, Nature 438, 17 Nov 2005.
2. Jamie Lloyd-Smith, UC Berkeley, Sebastian Schreiber, William &Mary College, Ekkehard Kopp, Hull, Wayne Getz, UC Berkeley: *The impact of individual variation in infectiousness on disease emergence*, Nature 438, 355-359, 17 Nov 2005