INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Hébert-Dufresne L, Althouse BM, Scarpino SV, Allard A. 2020 Beyond R_0: heterogeneity in secondary infections and probabilistic epidemic forecasting. *J. R. Soc. Interface* **17**: 20200393.
http://dx.doi.org/10.1098/rsif.2020.0393

Received: 25 May 2020
Accepted: 12 October 2020

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
computational biology

Keywords:
epidemiology, complex networks, branching processes

Author for correspondence:
Antoine Allard
e-mail: antoine.allard@phy.ulaval.ca

Beyond R_0: heterogeneity in secondary infections and probabilistic epidemic forecasting

Laurent Hébert-Dufresne[1,2,3], Benjamin M. Althouse[4,5,6], Samuel V. Scarpino[7,8,9,10,11,12] and Antoine Allard[3,13]

[1]Vermont Complex Systems Center, and [2]Department of Computer Science, University of Vermont, Burlington, VT 05405, USA
[3]Département de physique, de génie physique et d'optique, Université Laval, Québec, Canada G1V 0A6
[4]Institute for Disease Modeling, Bellevue, WA 98005, USA
[5]Information School, University of Washington, Seattle, WA 98195-2840, USA
[6]Department of Biology, New Mexico State University, Las Cruces, NM 88003, USA
[7]Network Science Institute, [8]Department of Marine and Environmental Sciences, [9]Department of Physics, and [10]Department of Health Sciences, Northeastern University, Boston, MA 02115, USA
[11]ISI Foundation, Turin 10126, Italy
[12]Santa Fe Institute, Santa Fe, NM 87501, USA
[13]Centre interdisciplinaire en modélisation mathématique, Université Laval, Québec, Canada G1V 0A6

BMA, 0000-0002-5464-654X; AA, 0000-0002-8208-9920

The basic reproductive number, R_0, is one of the most common and most commonly misapplied numbers in public health. Often used to compare outbreaks and forecast pandemic risk, this single number belies the complexity that different epidemics can exhibit, even when they have the same R_0. Here, we reformulate and extend a classic result from random network theory to forecast the size of an epidemic using estimates of the distribution of secondary infections, leveraging both its average R_0 and the underlying heterogeneity. Importantly, epidemics with lower R_0 can be larger if they spread more homogeneously (and are therefore more robust to stochastic fluctuations). We illustrate the potential of this approach using different real epidemics with known estimates for R_0, heterogeneity and epidemic size in the absence of significant intervention. Further, we discuss the different ways in which this framework can be implemented in the data-scarce reality of emerging pathogens. Lastly, we demonstrate that without data on the heterogeneity in secondary infections for emerging infectious diseases like COVID-19 the uncertainty in outbreak size ranges dramatically. Taken together, our work highlights the critical need for contact tracing during emerging infectious disease outbreaks and the need to look beyond R_0.

1. Introduction

In 1918, a typical individual infected with influenza transmitted the virus to between one and two of their social contacts [1], giving a value of the basic reproductive number—R_0, the expected number of secondary infections by a single infected individual introduced in a completely susceptible population—of between 1 and 2. These are similar to values of R_0 for the 2014 West Africa Ebola virus outbreak, yet Ebola virus disease infected a tenth of 1% of the number of individuals believed to have been infected by the 1918 influenza virus [2,3]. The two diseases are of course vastly different in symptoms and mortality, but most models to estimate the final size of an epidemic tend to ignore these features and instead focus on the actual spread through secondary infections. Similarly, the century separating the two epidemics saw vast improvements in healthcare and public health measures, as well as changes in human behaviour, which all help explain the massive discrepancy between Ebola virus disease in

THE ROYAL SOCIETY PUBLISHING

2014 and influenza in 1918 [4]. There is another critical but sometimes overlooked difference between these two diseases: heterogeneity in the number of secondary cases resulting from a single infected individual. Indeed, most individuals infected with Ebola virus gave rise to zero additional infections while a few gave rise to more than 10 [5,6]. Here, we demonstrate analytically that quantifying the variability in the number of secondary infections is critically important for quantifying the transmission risk of common and novel pathogens.

The basic reproduction number of an epidemic, R_0, is the expected number of secondary cases (note, we use the word 'case' in a generic sense to represent any infection, even if too mild to meet the clinical case definition [7]) produced by a primary case over the course of their infectious period in a completely susceptible population [8]. It is a simple metric that is commonly used to describe and compare the transmissibilty of emerging and endemic pathogens [9]. If $R_0 = 2$, one case turns to two, on average, and two turn to four as the epidemic grows. Conversely, the epidemic will die out if $R_0 < 1$.

Almost 100 years ago, work from Kermack & McKendrick [10–12] first demonstrated how to estimate the final size of an epidemic, integrating over all time to ignore the dynamics and focus on the final fraction of individuals reached by the epidemic, $R(\infty)$. Specifically, they considered a scenario such that:

(i) the disease results in complete immunity or death,
(ii) all individuals are equally susceptible,
(iii) the disease is transmitted in a closed population,
(iv) contacts occur according to the law of mass action, and
(v) the population is large enough to justify a deterministic analysis.

Under these assumptions, Kermack and McKendrick showed that an epidemic with a given R_0 will infect a fixed fraction $R(\infty)$ of the susceptible population by solving

$$R(\infty) = -\frac{1}{R_0}\ln[1 - R(\infty)]. \tag{1.1}$$

This solution describes a final outbreak size equal to 0 when $R_0 \leq 1$ and increasing roughly as $1 - \exp(-R_0)$ when $R_0 > 1$. Therefore, a larger R_0 leads to a larger outbreak, which infects the entire population in the limit $R_0 \to \infty$. This direct relationship between R_0 and the final epidemic size is at the core of the conventional wisdom that a larger R_0 will cause a larger outbreak. Unfortunately, the equation relating R_0 to final outbreak size from Kermack and McKendrick is only valid when all the above assumptions hold, which is rare in practice.

As a result, relying on R_0 alone is often misleading when comparing different pathogens or outbreaks of the same pathogen in different settings [13–15]. This is especially critical considering that many outbreaks are not shaped by the 'average' individuals but rather by a minority of super-spreading events [13,16,17]. To more fully quantify how heterogeneity in the number of secondary infections affects outbreak size, we turn towards network epidemiology and derive an equation for the total number of infected individuals using all moments of the distribution of secondary infections.

2. Random network analysis

Random network theory allows us to relax some of assumptions made by Kermack and McKendrick, mainly to account for heterogeneity and stochasticity in the number of secondary infections caused by a given individual. We first follow the analysis of [18] and define

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k \tag{2.1}$$

as the probability generating function (PGF) of the distribution $\{p_k\}$ of the number of contacts individuals have (their *degree*). In other words, a randomly chosen node has a degree equal to k with probability p_k. If we instead select an edge at random, the degree of the node at either of its two ends will be distributed according to $k\,p_k/\langle k\rangle$ since an edge is k times more likely to reach a node of degree k than a node of degree 1. Here $\langle k\rangle = \sum_{k=0}^{\infty} k p_k$ is the average degree and acts as a normalization constant. We define the *excess* degree as the number of *other* edges a node has when it has been reached via one of its edges. Since the excess degree equals the degree of a node at the end of an edge minus 1, the excess degree distribution is generated by

$$G_1(x) = \frac{1}{\langle k\rangle}\sum_{k=1}^{\infty} k p_k x^{k-1} = \frac{G_0'(x)}{G_0'(1)}, \tag{2.2}$$

where $G_0'(x)$ denotes the derivative of $G_0(x)$ with respect to x.

We now assume that the network in question is the network of all edges that *will* transmit a disease if either of the two nodes at its ends were infected. Consequently, $G_1(x)$ generates the number of secondary infections that individual nodes would cause if infected. Consequently, the connected component to which a node belongs (the maximal subset of nodes between which paths exist between all pairs of nodes) will be infected should that node be the first infected individual (the patient zero). In this framework, the size of the largest possible epidemic corresponds to the size of the giant connected component (GCC).

To calculate the size of the GCC, we first look for the probability u that following a random edge leads to a node *not* part of the GCC. For that node to not be a part of the GCC, none of its *other* neighbours should belong to it either, which occurs with probability u^{k-1} if that node has a degree equal to k. Since u is defined for any edge, we take the average over the excess degree distribution, which yields the self-consistent equation whose solution is u

$$u = \frac{1}{\langle k\rangle}\sum_{k=1}^{\infty} k p_k u^{k-1} = G_1(u). \tag{2.3}$$

Equation (2.3) is a condition of self-consistency since both sides describe the same quantity, u, under two different perspectives, which allows us to solve for u. The left-hand side is our definition of the probability u that a random edge followed in one direction does not lead to an infinite component; whereas the right-hand side calculates this probability from the perspective of the excess degree of the node reached through the random edge. The size of the GCC is a fraction of the full population N that we will denote $R(\infty)$ because it corresponds to the potential, macroscopic, outbreak size. Noting that a node of degree k has *no* edge leading to the GCC with probability u^k, $R(\infty)$ corresponds to the fraction of nodes with at least one edge leading to the GCC

$$R(\infty) = \sum_{k=0}^{\infty} p_k(1 - u^k) = 1 - G_0(u). \tag{2.4}$$

Data on the distribution of secondary infections inform us about $G_1(x)$ directly, but our choice of $G_0(x)$ represents our

assumptions on patient zero: is the first case different from subsequent cases? If not, we could use $G_0(x) = G_1(x)$ to obtain final size estimates of a branching process as described in [13] but that would ignore the fact that patient zero was not chosen by following a person-to-person transmission link, a network bias described in [19]. When assuming a relationship between $G_0(x)$ and $G_1(x)$ as in equation (2.2), $G_0(x)$ will still have one degree of freedom remaining, p_0, which requires further assumptions to be made to set its value (which we introduce in equation (3.8)). Putting all these different assumptions under the same framework will allow us to explicitly compare them.

Regardless of the specifics of the chosen model and of its underlying assumptions, equation (2.4) provides the size of the largest possible epidemic in the limit of infinite population size. Similarly to the Kermack–McKendrick solution, this approach provides an almost exact mapping to the final size of the dynamical spreading process without describing the temporal dynamics since we are effectively integrating over time by considering only transmissions that occur and ignoring when they occur [20]. There are however methods to use a branching process perspective or extend PGFs to temporal dynamics by considering inter-generation time [21,22].

3. Results

The network approach naturally accounts for heterogeneity, meaning that some individuals will cause more infections than others. The network approach also accounts for stochasticity explicitly: even with $R_0 > 1$, there is a probability $1 - R(\infty)$ that patient zero lies outside of the giant outbreak and therefore only leads to a small outbreak that does not invade the population. However, the analysis in terms of PGFs is obviously more involved than simply assuming mass-action mixing and solving equation (1.1). In fact, the PGFs $G_0(x)$ or $G_1(x)$ require a full distribution of secondary cases, which will in practice involve the specification of a high-order polynomial. Previous network models [19,23] tend to specify $G_0(x)$ then derive $G_1(x)$, but our approach focuses on secondary infections and $G_1(x)$ to unify the network and branching process perspectives [13,24]. Doing so clarifies our assumptions and allows us to simplify further.

To further this approach, we propose reformulating the classic network model in terms of the cumulant generating function (CGF) of secondary cases. The CGF $K(y)$ of a random variable X can be written as $K(y) = \sum \kappa_n y^n/n!$, where κ_n are the cumulants of the distribution of secondary infections. These are useful because the cumulants are easier to interpret, i.e. κ_1 is simply the average number of secondary cases R_0, κ_2 is the variance, κ_3 is related to the skewness and κ_4 is related to the kurtosis of the full distribution, etc. By definition, a PGF $G(x)$ of a random variable is linked to $K(y)$ through $G(x) = \exp[K(\ln x)]$. Therefore, we can replace the PGF $G_1(x)$ for the distribution of secondary infections by a function in terms of the cumulants of that distribution.

3.1. Analysis of cumulants and derivation of Kermack–McKendrick

We can easily derive Kermack and McKendrick's result from this framework since their solution assumes a well-mixed population, which corresponds to a Poisson distribution of secondary infections. We first re-write $G_1(x)$ in terms of the cumulants κ_n as

$$G_1(x) = \exp\left[\sum_{n=1}^{\infty} \frac{1}{n!} \kappa_n (\ln x)^n\right], \tag{3.1}$$

which is a particularly convenient representation for a Poisson distribution because its cumulants $\kappa_n = R_0$ for all $n > 0$. Moreover, since $G_0(x) = G_1(x)$ in the Poisson case, the final outbreak size of the Kermack–McKendrick analysis will be set by $u_{\mathrm{KM}} = G_1(u_{\mathrm{KM}})$, or

$$u_{\mathrm{KM}} = \exp\left[\sum_{n=1}^{\infty} \frac{1}{n!} R_0 (\ln u_{\mathrm{KM}})^n\right] = \exp\left[R_0(u_{\mathrm{KM}} - 1)\right]$$
$$\hookrightarrow R_{\mathrm{KM}}(\infty) = 1 - \exp[R_0(u_{\mathrm{KM}} - 1)] = 1 - \exp[-R_0 R_{\mathrm{KM}}(\infty)]. \tag{3.2}$$

Taking the logarithm of the exponential term from this last equation yields equation (1.1).

The solution to $u = G_1(u)$ gives the probability that every infection caused by patient zero fails to generate an epidemic. For more general distributions, it is useful to rewrite equation (3.1) as

$$u = G_1(u) = \exp\left[\sum_{n=1}^{\infty} \frac{1}{n!} \kappa_n (\ln u)^n\right]$$
$$= \exp\left[R_0 |\ln u| - \frac{1}{2}\sigma^2 |\ln u|^2 + \frac{1}{6}\kappa_3 |\ln u|^3 - \frac{1}{24}\kappa_4 |\ln u|^4 \ldots\right] \tag{3.3}$$

to highlight its alternating nature because the logarithm of u is negative (u is a probability) such that its nth power is positive when n is even and negative when n is odd.

The alternating sign of contribution from high-order moments in equation (3.3) can be interpreted as follows. A disease needs a high average number of secondary infections (high $\kappa_1 = R_0$) to spread, but, given that average, a disease with small variance in secondary infections will spread much more reliably and be less likely to stochastically die out. Given a variance, a disease with high skewness (i.e. with positive deviation contributing to most of the variance) will be more stable than a disease with negative skewness (i.e. with most deviations being towards small secondary infections). Given a skewness, a disease will be more stable if it has frequent small positive deviations rather than infrequent large deviations—hence a smaller kurtosis—as stochastic die out could easily occur before any of those large infrequent deviations occur.

Our re-interpretation already highlights a striking result: higher moments of the distribution of secondary cases can lead a disease with a lower R_0 to invade a population more easily and to reach a larger final outbreak size than a disease with a higher R_0. This result is illustrated in figure 1.

3.2. Normal distributions and the impact of variance

A second useful application of the cumulants formulation involves diseases with a large reproductive number R_0 whose distribution of secondary infections can be convincingly modelled by a normal distribution. Using a normal distribution for the distribution of secondary infections is only valid for very large R_0 since we have to both model a discrete distribution with a continuous one and ignore

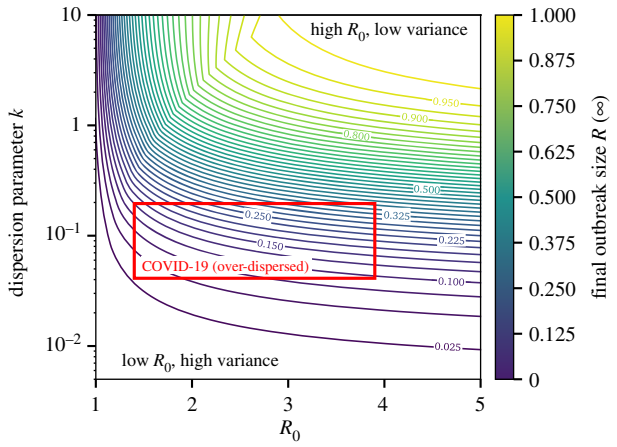

Figure 1. Final size of outbreaks with different average R_0 and heterogeneity k in the distribution of secondary cases. We use a negative binomial distribution of secondary cases and scan a realistic range of parameters. The range of parameters corresponding to estimates for COVID-19 based on a binomial negative distribution in large populations is highlighted by a red box (see [25] and table 1). Most importantly, with fixed average, the dispersion parameter is inversely proportional to the variance of the underlying distribution of secondary cases. The degree of freedom, p_0, is here set by setting the average number of infections around patient zero to be less than or equal to R_0. The Kermack–McKendrick solution would correspond to the limit $k \to \infty$, and could be more appropriate in some dense and well-mixed settings.

negative numbers of secondary infections. The advantage of this approximation is that while the raw moments of a normal distribution are quite complicated, the cumulants are simple: κ_1 is equal to the mean R_0, κ_2 is equal to the variance σ^2 and all other cumulants are 0. We can thus write

$$G_1(x) = \exp\left[R_0 \ln x + \frac{1}{2}\sigma^2 (\ln x)^2\right] = x^{R_0 + \frac{\sigma^2}{2}\ln x} \tag{3.4}$$

and solving for $u = G_1(u)$ yields

$$u = \exp\left[-\frac{2}{\sigma^2}(R_0 - 1)\right]. \tag{3.5}$$

This equation can then be used for direct comparison of the probability of invasion of two different diseases with normal distributions of secondary infections. Given a transmission event from patient zero to a susceptible individual, disease B will be more likely to invade the population than disease A if

$$\frac{\sigma_A^2}{\sigma_B^2} < \frac{R_{0,A} - 1}{R_{0,B} - 1}. \tag{3.6}$$

For example, a disease with half the basic reproductive number of another will still be more likely to invade a population and lead to a larger outbreak if its variance is less than or close to half the variance of the other disease.

Altogether, the results of the previous subsections show that taking into account the contribution of these higher moments should yield different, hopefully better, estimates for the final size of real outbreaks. To test this hypothesis, we now introduce a more specific network model.

3.3. Negative binomial network model

We present a specific network model assuming the number of secondary infections to be distributed according to a negative binomial distribution parametrized by its average R_0 and dispersion k [13]. Its PGF is

$$G_1(x) = \sum_{n=0}^{\infty} n + k - 1 n \left[\frac{R_0}{R_0 + k}\right]^n \left[1 - \frac{R_0}{R_0 + k}\right]^k x^n = \left[1 + \frac{R_0}{k}(1 - x)\right]^{-k}. \tag{3.7}$$

The general network theory formalism requires the specification of the PGF $G_0(x)$ that is related to $G_1(x)$ via equation (2.2). Specifying $G_1(x)$ therefore fixes $G_0(x)$ up to a constant and to a multiplicative factor. Without loss of generality, we set

$$G_0(x) = p_0 + (1 - p_0) g_0(x) \tag{3.8}$$

with $0 \leq p_0 \leq 1$, $g_0(0) = 0$ and $g_1(1) = 1$. Equation (2.2) becomes

$$G_1(x) = \frac{g_0'(x)}{g_0'(1)}, \tag{3.9}$$

from which we compute

$$g_0(x) = \int g_0'(x)\,\mathrm{d}x = g_0'(1) \int G_1(x)\,\mathrm{d}x = -\frac{k g_0'(1)}{R_0(1 - k)}\left[1 + \frac{R_0}{k}(1 - x)\right]^{1-k} + C, \tag{3.10}$$

with $k \neq 1$, and where C and $g_0'(1)$ are fixed by imposing $g_0(0) = 0$ and $g_1(1) = 1$. Rearranging the terms, we find that

$$g_0(x) = \frac{1 - \left[1 - \frac{R_0 x}{R_0 + k}\right]^{1-k}}{1 - \left[\frac{k}{R_0 + k}\right]^{1-k}}, \tag{3.11}$$

from which we finally obtain

$$G_0(x) = p_0 + (1 - p_0)\frac{1 - \left[1 - \frac{R_0 x}{R_0 + k}\right]^{1-k}}{1 - \left[\frac{k}{R_0 + k}\right]^{1-k}} \tag{3.12}$$

with $k \neq 1$. The case $k = 1$ must be treated separately and yields

$$G_0(x) = p_0 + (1 - p_0)\left[1 - \frac{\ln[1 + R_0(1 - x)]}{\ln[1 + R_0]}\right]. \tag{3.13}$$

From equations (3.12) and (3.13), we find that the average number of secondary infections caused by patient zero is

$$G_0'(1) = (1 - p_0)\frac{(1 - k)R_0}{k}\frac{1}{\left[\frac{k}{R_0 + k}\right]^{k-1} - 1} \tag{3.14}$$

if $k \neq 1$, and

$$G_0'(1) = (1 - p_0)\frac{R_0}{\ln[1 + R_0]} \tag{3.15}$$

if $k = 1$. The average number of secondary infections caused by patient zero can therefore be greater or smaller than R_0. Since patient zero should not be expected to create *more* secondary cases than the next generation of infections, we set the value of $p_0 \in [0, 1]$ such that $G_0'(1)$ is as close as possible to R_0 whenever $G_0'(1) > R_0$.

Table 1. Estimates for R_0 and for the negative binomial distribution dispersion parameter, k, used in figure 2 ([a] and [b], respectively, denote 95% and 90% confidence intervals). The proportion of susceptible individuals infected as reported either in the literature or by the US Centers for Disease Control and Prevention. For severe acute respiratory syndrome (SARS) the proportion of infected was taken from serosurveys among wild animal handlers (15%) and among healthcare workers (<1%) [27]. For influenza (2009), we took data on school-aged children. For COVID-19, we present emerging evidence surrounding the final proportion of infected individuals after the first outbreak waves at the level of large communities [28,29] and a school [30], which all fall around 15%, and at the level of dense groups like a fishing vessel with a value around 86% [31]. Note that the estimates of the proportion of infected individuals, for R_0 and for k, were not necessarily inferred from the same populations. Such information is rarely, if ever, available for the same outbreak, unfortunately. COVID-19, coronavirus disease 2019; MERS, Middle East respiratory syndrome.

disease	location	year	prop. infect.	R_0	k	reference
MERS	global	2013	0%	0.47 (0.29–0.80)[a]	0.26 (0.09–1.24)[a]	[21,32]
SARS	global	2003	0–15%	1.63 (0.54–2.65)[b]	0.16 (0.11–0.64)[b]	[13,27,33]
smallpox	Europe	1958–1973	55%	3.19 (1.66–4.62)[b]	0.37 (0.26–0.69)[b]	[13,34]
influenza	Baltimore (USA)	1918	40%	1.77 (1.61–1.95)[a]	0.94 (0.59–1.72)[a]	[35,36]
influenza	Italy	2009	39%	1.321 (1.299–1.343)[a]	8.092 (5.170–11.794)[a]	[37,38]
COVID-19	global	2020	13–16% and 86%	2.5 (1.4–12)[a]	0.1 (0.04–1)[a]	[25,28–31,39–41]

A large-scale epidemic is predicted by this framework [18] if

$$G_1'(1) = R_0 > 1, \tag{3.16}$$

as in the analysis by Kermack & McKendrick [10–12]. Its size, $R(\infty)$, is computed with $G_0(x)$ as

$$R(\infty) = 1 - G_0(u), \tag{3.17}$$

where u is the solution of

$$u = G_1(u), \tag{3.18}$$

which we solve using the relaxation method [26] with an initial condition randomly chosen in the open interval (0, 1).

3.4. Comparison of estimators with empirical data

We now compare the final outbreak size estimates from equation (1.1) (Kermack and McKendrick) with estimates from equation (3.17) with a negative binomial offspring distribution (table 1). Ideally, this validation would use estimates of final outbreak size, R_0 and k inferred from the same population, but unfortunately these are rarely, if ever, available. Similarly, once interventions are put in place and/or substantial behavioural change occurs, all methods that do not account for these effects will over-estimate the total outbreak size [42]. To attenuate some of these issues, we focus on outbreaks where no vaccine was available or before large interventions were put in place: smallpox in unvaccinated populations, the 1918 influenza pandemic, school children prior to the availability of the 2009 H1N1 vaccine, as well as for severe acute respiratory syndrome (SARS) among specific communities such as wild animal handlers (other smaller estimates correspond to healthcare workers). Importantly, focusing on smaller local outbreaks also allows us to mitigate any effect of reseeding in the same population as our approach describes a single transmission chain.

As predicted, figure 2 shows that the Kermack and McKendrick formulation consistently and significantly over-predicts the outbreak size across six different pathogens where we could find confidence interval estimates for R_0 and for the negative binomial over-dispersion parameter (k). All network approaches produce estimates of the total outbreak size which are consistent with reported prevalence. Despite the inherent problems associated with such validations, network models appear to provide a much more reasoned estimate of the total risk to any given population, and predictions very close to the most recent seropositivity estimates for the COVID-19 outbreak in a German municipality [28] and in obstetric patients presenting for delivery [29].

4. Discussion

From re-emerging pathogens like yellow fever and measles to emerging threats like Middle East respiratory syndrome coronavirus and Ebola, the World Health Organization monitored 119 different infectious disease outbreaks in 2019 alone [43]. For each of these outbreaks, predicting both the epidemic potential and the most likely number of cases is critically important for efficient and effective responses. This need for rapid situational awareness is why R_0 is so widely used in public health. However, our main analysis shows that not only is R_0 insufficient in fully determining the final size of an outbreak, but having a larger outbreak with a lower R_0 is relatively easy considering the randomness associated with most transmission events and the heterogeneity of physical contacts. To address the need for rapid quantification of risk, while acknowledging the shortcomings of R_0, we use network science methods to derive both the probability of an epidemic and its final size.

These results are not without important caveats. Specifically, we must remember that distributions of secondary cases, just like R_0 itself, are just as much a product of a pathogen as of the population in which it spreads. For example, aspects of the social contact network [44], metapopulation structure [45], human mobility [46], adaptive behaviour [47] and even other pathogens [48,49] all interact to cause complex patterns of disease emergence, spread and persistence. Therefore, great care must be taken when using any of these tools to compare outbreaks or to inform current events with past data. In addition, it remains a challenge to determine the final outbreak size in the absence of interventions, re-seeding, etc., and after properly accounting for the initial number of infectious individuals and the proportion of the population that is susceptible to infection. For these reasons, we focused on empirical studies that included data on the initial conditions in the population.

Figure 2 only used a few known outbreaks to validate the different approaches because data on secondary cases are

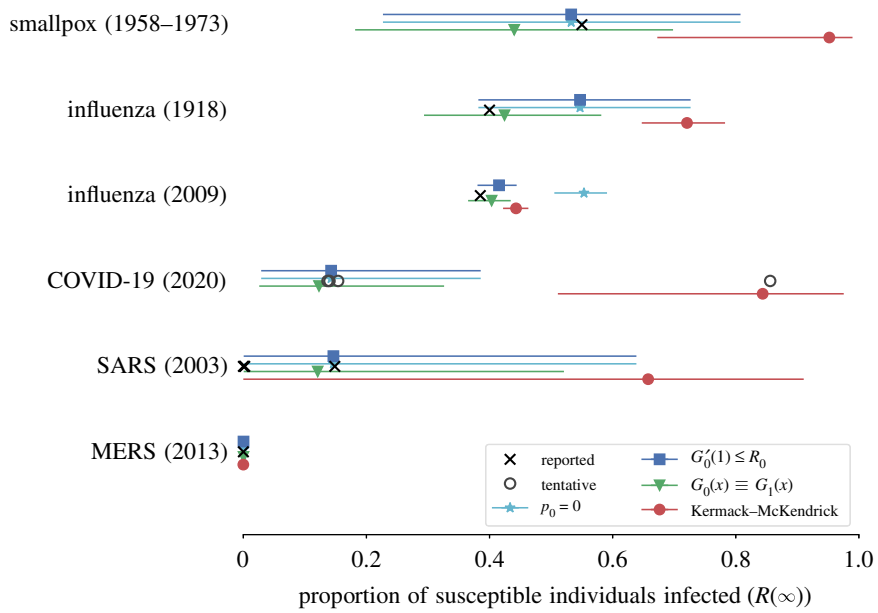

Figure 2. Using published estimates of R_0 and the dispersion parameter k, we estimated the total outbreak size for six different diseases using three versions of the network approach and compared them with the classic Kermack–McKendrick solution. The confidence intervals span the range of uncertainty reported for R_0 and k. The black markers show reported total outbreak sizes (total proportion of susceptible individuals infected) for each disease. For influenza, we report the estimated proportion of school-aged children infected. For COVID-19, we use tentative markers showing the range of attack rates measured in different contexts as there is currently no consensus for what constitutes a typical COVID-19 outbreak. We highlight though the differences between the final size estimates for COVID-19: most typify the observed over-dispersed nature of transmission, except for the outbreak on a fishing vessel (right side point) where contacts are more well mixed and thus better characterized by a Kermack–McKendrick transmission process. The red circles are the estimated proportion infected using the method developed by Kermack and McKendrick, i.e. equation (1.1). The other markers show the estimated proportion infected obtained with equation (3.17) under different assumptions about patient zero: the model described in the main text, which ensures that the expected number of secondary infections caused by patient zero is at most R_0 (blue squares); the same model but assuming $p_0 = 0$ such that no individuals have exactly zero contact (cyan stars); and a network version of [13], where $G_0(x) \equiv G_1(x)$ such that patient zero is no different from subsequent patients (green triangles). See table 1 for data and additional information.

rare. In practice, three types of data could potentially be used in real time to improve predictions by considering secondary case heterogeneity. First, contact tracing data, whose objective is to identify people who may have come into contact with an infectious individual. While mostly a preventive measure to identify cases before complications, it directly informs us about potential secondary cases caused by a single individual, and therefore provides us with an estimate for $G_1(x)$. Both for generating accurate predictions of epidemic risk and controlling the outbreak, it is vital to begin contact tracing before numerous transmission chains become widely distributed across space [50,51].

Second, viral genome sequences provide information on both the timing of the outbreak [52] and the structure of secondary cases [53]. For example, methods exist to reconstruct transmission trees for sampled sequences using simple mutational models to construct a likelihood for a specific transmission tree [54,55] and translate coalescent rates into key epidemiological parameters [56,57]. Despite the potential for genome sequencing to revolutionize outbreak response, the global public health community often struggled to coordinate data sharing across international borders, between academic researchers and with private companies [58–60]. However, the current COVID-19 pandemic has stimulated prompt and widespread sharing of genomic data; this will hopefully become standard in the future.

Third, early incidence data can be leveraged to infer parameters of the secondary case distribution through comparison with simulations. Comparing the output of agent-based simulations with reported incidence can be used to effectively sample a joint posterior distribution over R_0 and dispersion parameter k. This approach was used by most studies referenced in table 1. Most importantly, these simulations need not be run over long periods of time to predict final outbreak size. Instead, they only need to be run over enough early data to infer the parameter estimates that are then fed into our network model to compute the final outbreak size.

As for COVID-19, figure 1 shows how the width of the confidence interval on our prediction for the final outbreak size mostly stems from uncertainty in the heterogeneity of secondary infections, i.e. the dispersion parameter k. Note that the estimates for R_0 and k used here are from population-level estimates (table 1) and are therefore not representative of COVID-19 in all contexts. With limited heterogeneity, our predictions would have been closer to classic mass-action forecasts and the current pandemic of COVID-19 would probably have been a consequence not only of R_0 but also of the homogeneity of secondary infections: each new case steadily leading to additional infections. However, we note that emerging evidence, taken from a serosurvey in the municipality of Gangelt, Germany [28], and from universal testing in all obstetric patients presenting for delivery at two hospitals [29], suggests that the final size for a single, established COVID-19 transmission chain is around 15% of the population, which is both in agreement with estimates from our approach and far below the final size predicted by the Kermack and McKendrick formulation. With recent large estimates for its heterogeneity, the observed transmission could be mostly maintained by so-called 'superspreading events', which could be easier to manage with contact tracing, screening and infection control [61,62].

In conclusion, we reiterate that, when accounting for the full distribution of secondary cases caused by an infected individual, there is no direct relationship between R_0 and the size of an outbreak. We also stress that both R_0 and the full secondary case distribution are not properties of the disease itself, but are instead set by properties of the pathogen, the host population and the context of the outbreak. This is best exemplified by the widely different attack rates of COVID-19 observed in figure 2 between the fishing vessel (85.6%) and the school (13.7%). Both populations were roughly of the same size but contacts in the former are denser and much more homogeneously mixed, leading to an outbreak consistent with the Kermack–McKendrick solution while contacts in the latter follow heterogeneous classroom and age patterns leading to a lower outbreak size. Our methodology can straightforwardly translate any of these estimates of transmission heterogeneity into epidemic forecasts. Altogether, predicting outbreak size based on early data is an incredibly complex challenge but one that is increasingly within reach owing to new mathematical analyses and faster communication of public health data.

Data accessibility. See https://github.com/Emergent-Epidemics/beyond_R0 for data and scripts.

Authors' contributions. All authors designed and performed the research and wrote the manuscript.

Competing interests. We declare we have no competing interests.

Funding. L.H.-D. acknowledges support from the National Institutes of Health 1P20 GM125498-01 Centers of Biomedical Research Excellence Award. B.M.A. is supported by Bill and Melinda Gates through the Global Good Fund. S.V.S. is supported by start-up funds provided by Northeastern University. A.A. acknowledges financial support from the Sentinelle Nord initiative of the Canada First Research Excellence Fund and from the Natural Sciences and Engineering Research Council of Canada (project 2019-05183). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. Biggerstaff M, Cauchemez S, Reed C, Gambhir M, Finelli L. 2014 Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC Infect. Dis.* **14**, 480. (doi:10.1186/1471-2334-14-480)
2. Mills CE, Robins JM, Lipsitch M. 2004 Transmissibility of 1918 pandemic influenza. *Nature* **432**, 904–906. (doi:10.1038/nature03063)
3. Kaner J, Schaack S. 2016 Understanding Ebola: the 2014 epidemic. *Global. Health* **12**, 53. (doi:10.1186/s12992-016-0194-4)
4. Bootsma MCJ, Ferguson NM. 2007 The effect of public health measures on the 1918 influenza pandemic in U.S. cities. *Proc. Natl Acad. Sci. USA* **104**, 7588–7593. (doi:10.1073/pnas.0611071104)
5. WHO Ebola Response Team. 2014 Ebola virus disease in West Africa – the first 9 months of the epidemic and forward projections. *N. Engl. J. Med.* **371**, 1481–1495. (doi:10.1056/NEJMoa1411100)
6. Althaus CL. 2014 Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. *PLOS Curr.* (doi:10.1371/currents.outbreaks.91afb5e0f279e7f29e7056095255b288)
7. Althouse BM, Scarpino SV. 2015 Asymptomatic transmission and the resurgence of Bordetella pertussis. *BMC Med.* **13**, 146. (doi:10.1186/s12916-015-0382-8)
8. Diekmann O, Metz JAJ, Heesterbeek JAP. 1995 The legacy of Kermack and McKendrick. In *Epidemic models: their structure and relation to data* (ed. D Mollison), pp. 95–115. Cambridge, UK: Cambridge University Press.
9. Sheikh K, Watkins D, Wu J, Gröndahl M. 2020 How bad will the coronavirus outbreak get? Here are 6 key factors. *The New York Times*. See https://www.nytimes.com/interactive/2020/world/asia/china-coronavirus-contain.html.
10. Kermack WO, McKendrick AG. 1927 A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A* **115**, 700–721. (doi:10.1098/rspa.1927.0118)
11. Kermack WO, McKendrick AG. 1932 Contributions to the mathematical theory of epidemics. II. The problem of endemicity. *Proc. R. Soc. Lond. A* **138**, 55–83. (doi:10.1098/rspa.1932.0171)
12. Kermack WO, McKendrick AG. 1933 Contributions to the mathematical theory of epidemics. III. Further studies of the problem of endemicity. *Proc. R. Soc. Lond. A* **141**, 94–122. (doi:10.1098/rspa.1933.0106)
13. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. 2005 Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359. (doi:10.1038/nature04153)
14. Bansal S, Grenfell BT, Meyers LA. 2007 When individual behaviour matters: homogeneous and network models in epidemiology. *J. R. Soc. Interface* **4**, 879–891. (doi:10.1098/rsif.2007.1100)
15. Vizi Z, Kiss IZ, Miller JC, Röst G. 2019 A monotonic relationship between the variability of the infectious period and final size in pairwise epidemic modelling. *J. Math. Ind.* **9**, 1. (doi:10.1186/s13362-019-0058-7)
16. Meyers LA, Pourbohloul B, Newman MEJ, Skowronski DM, Brunham RC. 2005 Network theory and SARS: predicting outbreak diversity. *J. Theor. Biol.* **232**, 71–81. (doi:10.1016/j.jtbi.2004.07.026)
17. Althouse BM, Wenger EA, Miller JC, Scarpino SV, Allard A, Hébert-Dufresne L, Hu H. 2020 Stochasticity and heterogeneity in the transmission dynamics of SARS-CoV-2. (http://arxiv.org/abs/2005.13689).
18. Newman MEJ, Strogatz SH, Watts DJ. 2001 Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118. (doi:10.1103/PhysRevE.64.026118)
19. Newman MEJ. 2002 Spread of epidemic disease on networks. *Phys. Rev. E* **66**, 016128. (doi:10.1103/PhysRevE.66.016128)
20. Kenah E, Robins JM. 2007 Second look at the spread of epidemics on networks. *Phys. Rev. E* **76**, 036113. (doi:10.1103/PhysRevE.76.036113)
21. Kucharski AJ, Althaus CL. 2015 The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. *Eurosurveillance* **20**, 21167. (doi:10.2807/1560-7917.ES2015.20.25.21167)
22. Noël P-A, Davoudi B, Brunham RC, Dubé LJ, Pourbohloul B. 2009 Time evolution of epidemic disease on finite and infinite networks. *Phys. Rev. E* **79**, 026101. (doi:10.1103/physreve.79.026101)
23. Miller JC. 2018 A primer on the use of probability generating functions in infectious disease modeling. *Infect. Dis. Model.* **3**, 192–248. (doi:10.1016/j.idm.2018.08.001)
24. Nishiura H, Yan P, Sleeman CK, Mode CJ. 2012 Estimating the transmission potential of supercritical processes based on the final size distribution of minor outbreaks. *J. Theor. Biol.* **294**, 48–55. (doi:10.1016/j.jtbi.2011.10.039)
25. Endo A, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Abbott S, Kucharski AJ, Funk S. 2020 Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5**, 67.
26. Newman MEJ. 2012 *Computational physics*, p. 562. CreateSpace Independent Publishing Platform.
27. Leung GM *et al.* 2006 Seroprevalence of IgG antibodies to SARS-coronavirus in asymptomatic or subclinical population groups. *Epidemiol. Infect.* **134**, 211–221. (doi:10.1017/S0950268805004826)
28. Streeck H, Hartmann G, Exner M, Schmid M. 2020 Preliminary result and conclusions of the COVID-19 case cluster study (Gangelt Municipality). See https://www.land.nrw/sites/default/files/asset/document/zwischenergebnis_covid19_case_study_gangelt_en.pdf.
29. Sutton D, Fuchs K, D'Alton M, Goffman D. 2020 Universal screening for SARS-CoV-2 in women admitted for delivery. *N. Engl. J. Med.* **382**, 2163–2164. (doi:10.1056/NEJMc2009316)

30. Stein-Zamir C, Abramson N, Shoob H, Libal E, Bitan M, Cardash T, Cayam R, Miskin I. 2020 A large COVID-19 outbreak in a high school 10 days after schools' reopening, Israel, May 2020. *Eurosurveillance* **25**, 2001352. (doi:10.2807/1560-7917.ES.2020.25.29.2001352)
31. Addetia A, Crawford KHD, Dingens A, Zhu H, Roychoudhury P, Huang M-L, Jerome KR, Bloom JD, Greninger AL. 2020 Neutralizing antibodies correlate with protection from SARS-CoV-2 in humans during a fishery vessel outbreak with high attack rate. *J. Clin. Microbiol.* **58**, e02107-20. (doi:10.1128/JCM.02107-20)
32. Memish ZA *et al.* 2014 Prevalence of MERS-CoV nasal carriage and compliance with the Saudi health recommendations among pilgrims attending the 2013 Hajj. *J. Infect. Dis.* **210**, 1067–1072. (doi:10.1093/infdis/jiu150)
33. Quah SR, Hin-Peng L. 2004 Crisis prevention and management during SARS outbreak, Singapore. *Emerg. Infect. Dis.* **10**, 364–368. (doi:10.3201/eid1002.030418)
34. Mack TM, Thoma DB, Ali A, Khan MM. 1972 Epidemiology of smallpox in West Pakistan. *Am. J. Epidemiol.* **95**, 157–168. (doi:10.1093/oxfordjournals.aje.a121380)
35. Taubenberger JK, Morens DM. 2006 1918 Influenza: the mother of all pandemics. *Emerg. Infect. Dis.* **12**, 15–22. (doi:10.3201/eid1209.05-0979)
36. Fraser C, Cummings DAT, Klinkenberg D, Burke DS, Ferguson NM. 2011 Influenza transmission in households during the 1918 pandemic. *Am. J. Epidemiol.* **174**, 505–514. (doi:10.1093/aje/kwr122)
37. Kelly H, Peck HA, Laurie KL, Wu P, Nishiura H, Cowling BJ. 2011 The age-specific cumulative incidence of infection with pandemic influenza H1N1 2009 was similar in various countries prior to vaccination. *PLoS ONE* **6**, e21828. (doi:10.1371/journal.pone.0021828)
38. Dorigatti I, Cauchemez S, Pugliese A, Ferguson NM. 2012 A new approach to characterising infectious disease transmission dynamics from sentinel surveillance: application to the Italian 2009–2010 A/H1N1 influenza pandemic. *Epidemics* **4**, 9–21. (doi:10.1016/j.epidem.2011.11.001)
39. Li Q *et al.* 2020 Early transmission dynamics in Wuhan, China, of Novel Coronavirus–Infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207. (doi:10.1056/NEJMoa2001316)
40. Zhang Y, Li Y, Wang L, Li M, Zhou X. 2020 Transmission heterogeneity and super-spreading event of COVID-19 in a metropolis of China. *Int. J. Environ. Res. Public Health* **17**, 3705. (doi:10.3390/ijerph17103705)
41. Bi Q *et al.* 2020 Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *Lancet Infect. Dis.* **20**, 911–919. (doi:10.1016/S1473-3099(20)30287-5)
42. Eksin C, Paarporn K, Weitz JS. 2019 Systematic biases in disease forecasting—the role of behavior change. *Epidemics* **27**, 96–105. (doi:10.1016/j.epidem.2019.02.004)
43. WHO disease outbreaks by year: 2019. See https://www.who.int/csr/don/archive/year/2019/en (accessed 9 Februrary 2020).
44. Moreno Y, Pastor-Satorras R, Vespignani A. 2002 Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. B* **26**, 521–529. (doi:10.1140/epjb/e20020122)
45. Colizza V, Vespignani A. 2008 Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: theory and simulations. *J. Theor. Biol.* **251**, 450–467. (doi:10.1016/j.jtbi.2007.11.028)
46. Wesolowski A, Qureshi T, Boni MF, Sundsøy PR, Johansson MA, Rasheed SB, Engø-Monsen K, Buckee CO. 2015 Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proc. Natl Acad. Sci. USA* **112**, 11 887–11 892. (doi:10.1073/pnas.1504964112)
47. Scarpino SV, Allard A, Hébert-Dufresne L. 2016 The effect of a prudent adaptive behaviour on disease transmission. *Nat. Phys.* **12**, 1042–1046. (doi:10.1038/nphys3832)
48. Hébert-Dufresne L, Althouse BM. 2015 Complex dynamics of synergistic coinfections on realistically clustered networks. *Proc. Natl Acad. Sci. USA* **112**, 10 551–10 556. (doi:10.1073/pnas.1507820112)
49. Hébert-Dufresne L, Scarpino SV, Young J-G. 2020 Macroscopic patterns of interacting contagions are indistinguishable from social reinforcement. *Nat. Phys.* **16**, 426–431. (doi:10.1038/s41567-020-0791-2)
50. Dhillon RS, Srikrishna D. 2018 When is contact tracing not enough to stop an outbreak?. *Lancet Infect. Dis.* **18**, 1302–1304. (doi:10.1016/S1473-3099(18)30656-X)
51. Klinkenberg D, Fraser C, Heesterbeek H. 2006 The effectiveness of contact tracing in emerging epidemics. *PLoS ONE* **1**, e12. (doi:10.1371/journal.pone.0000012)
52. Smith GJD *et al.* 2009 Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–1125. (doi:10.1038/nature08182)
53. Scarpino SV *et al.* 2015 Epidemiological and viral genomic sequence analysis of the 2014 Ebola outbreak reveals clustered transmission. *Clin. Infect. Dis.* **60**, 1079–1082. (doi:10.1093/cid/ciu1131)
54. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. 2014 Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLOS Comput. Biol.* **10**, e1003457. (doi:10.1371/journal.pcbi.1003457)
55. Campbell F, Cori A, Ferguson N, Jombart T. 2019 Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLOS Comput. Biol.* **15**, e1006930. (doi:10.1371/journal.pcbi.1006930)
56. Volz EM, Koelle K, Bedford T. 2013 Viral phylodynamics. *PLOS Comput. Biol.* **9**, e1002947. (doi:10.1371/journal.pcbi.1002947)
57. Bouckaert R *et al.* 2019 BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650. (doi:10.1371/journal.pcbi.1006650)
58. Gardy J, Loman NJ, Rambaut A. 2015 Real-time digital pathogen surveillance—the time is now. *Genome Biol.* **16**, 155. (doi:10.1186/s13059-015-0726-x)
59. Van Puyvelde S, Argimon S. 2019 Sequencing in the time of Ebola. *Nat. Rev. Microbiol.* **17**, 5. (doi:10.1038/s41579-018-0130-0)
60. Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, Andersen KG. 2019 Tracking virus outbreaks in the twenty-first century. *Nat. Microbiol.* **4**, 10–19. (doi:10.1038/s41564-018-0296-2)
61. Hellewell J *et al.* 2020 Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Glob. Health* **8**, e488–e496. (doi:10.1016/S2214-109X(20)30074-7)
62. Kojaku S, Hébert-Dufresne L, Ahn Y-Y. 2020 The effectiveness of contact tracing in heterogeneous networks. (http://arxiv.org/abs/2005.02362).