RESEARCH ARTICLE

Characterizing the Transmission Potential of Zoonotic Infections from Minor Outbreaks

Adam J. Kucharski[1,2]*, W. John Edmunds[1]

1 Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom, **2** Fogarty International Center, National Institutes of Health, Bethesda, Maryland, United States

* adam.kucharski@lshtm.ac.uk

Abstract

The transmission potential of a novel infection depends on both the inherent transmissibility of a pathogen, and the level of susceptibility in the host population. However, distinguishing between these pathogen- and population-specific properties typically requires detailed serological studies, which are rarely available in the early stages of an outbreak. Using a simple transmission model that incorporates age-stratified social mixing patterns, we present a novel method for characterizing the transmission potential of subcritical infections, which have effective reproduction number R<1, from readily available data on the size of outbreaks. We show that the model can identify the extent to which outbreaks are driven by inherent pathogen transmissibility and pre-existing population immunity, and can generate unbiased estimates of the effective reproduction number. Applying the method to real-life infections, we obtained accurate estimates for the degree of age-specific immunity against monkeypox, influenza A(H5N1) and A(H7N9), and refined existing estimates of the reproduction number. Our results also suggest minimal pre-existing immunity to MERS-CoV in humans. The approach we describe can therefore provide crucial information about novel infections before serological surveys and other detailed analyses are available. The methods would also be applicable to data stratified by factors such as profession or location, which would make it possible to measure the transmission potential of emerging infections in a wide range of settings.

Author Summary

The transmission potential of a new infection depends on both the transmissibility of the pathogen and the level of immunity in the host population. However, it can be difficult to measure these properties if there are limited experimental studies of population immunity. By incorporating social contact patterns into a mathematical model of disease transmission, we show that it is possible to estimate both pathogen transmissibility and pre-existing immunity from available data on the size of outbreaks. When an infection does not transmit efficiently between humans, estimates often have to be made using case data from a limited number of small outbreaks. We find that, even with limited data, our technique

> can accurately evaluate the transmission potential of 'stuttering' chains of infection. We use the method to characterise transmission of four real infections: monkeypox, influenza A(H5N1) and A(H7N9) and MERS-CoV.

Introduction

Infections that spill over into humans from an external reservoir have the potential to cause epidemics with substantial morbidity and mortality, particularly if there is limited pre-existing immunity in the host population [1, 2]. However, novel pathogens do not always transmit efficiently when first introduced into human populations. Outbreaks of infections such as Middle East respiratory syndrome coronavirus (MERS-CoV) [3, 4] and monkeypox [5] have generally occurred as 'stuttering chains' of transmission [6], generating a relatively small number of linked clusters of cases without evidence of sustained transmission. Infections such as influenza A(H5N1) [7] and A(H7N9) [8] also appear to be subcritical at present, having so far failed to transmit efficiently between humans.

To assess the risk posed by novel infections, it is important to quantify their transmission potential. Transmissibility can be summarised using the effective reproduction number, R, defined as the average number of secondary cases produced by a typical infectious host [9]. The reproduction number can be separated into two components: the inherent transmissibility of a pathogen, and the level of susceptibility in the host population. In some circumstances, susceptibility might be reduced as a result of pre-existing immunity from previous vaccination campaigns, as is the case with monkeypox [5, 10], or prior exposure to a similar pathogen, as has been suggested for influenza A/H1N1p [11]. Such immunity will not necessary be distributed evenly across the population: if pathogens circulate over an extended period of time, or vaccination campaigns have been discontinued, pre-existing immunity is more likely to be found in older age groups [12].

It has been shown that the size distribution of minor outbreaks can provide information about the value of the effective reproduction number [13, 14]. However, existing techniques for estimating transmission potential from outbreak size data generally represent transmission in the host population using single-type branching process [15, 16, 17, 18]. As a result, it is not possible to distinguish between inherent pathogen transmissibility and population susceptibility. For instance, a highly transmissible pathogen in a mostly immune population might have the same effective reproduction number as an infection with lower inherent transmissibility spreading between fully susceptible hosts.

To characterize both the inherent properties of the pathogen and the level of population immunity, we analysed both the size and age distribution of minor outbreaks. Individuals of different ages have heterogeneous social contact patterns and hence different risks of infection during an outbreak [19, 20, 21]. Pre-existing immunity in older age groups can alter this pattern [22], making it possible to separate the reproduction number into its pathogen- and population-specific components. We made use of this observation by developing a novel age-structured model of stuttering transmission chains, which combined reported social contact data with a multi-type branching process [23, 24].

First we derived an expression for the outbreak size distribution in an age-stratified population, in which transmission between different age groups depended on the number of physical contacts reported in the POLYMOD survey in Great Britain. Next, we used simulated outbreaks to examine whether the model could distinguish between different types of infection using only age-stratified final outbreak size data. Finally, we analysed observed outbreak data

for monkeypox, influenza A(H5N1), A(H7N9) and MERS-CoV, and found that it was possible to accurately characterize pathogen transmissibility and pre-existing host immunity.

Results

Outbreak size distributions for age-structured populations

We explored the age pattern of infection by calculating the joint outbreak size distribution across different age groups. It has been suggested that the post-childhood drop in risky contacts that occurs around age 20 is a dominant factor shaping influenza dynamics [25], and the intense contacts between children make them an important epidemiological group for respiratory infections [26, 12]. We therefore divided the population into two groups: under 20 and over 20 year olds.

In a homogeneously mixing population, all individuals generated the same mean number secondary cases in the model (Fig. 1A). When the infection was introduced into the under 20 age group, the outbreak size distribution was therefore relatively symmetric between the two groups (Fig. 1B). When the offspring distribution of secondary cases depended on reported physical contacts between different groups in the UK (S1A Fig.), this pattern changed. Each infected host could generate secondary cases in either group, and the mean number of cases generated depended on which group the infected host was in (Fig. 1C). We assumed a fully susceptible population, which meant that the average number of secondary cases generated by a typical infectious individual was equal to the basic reproduction number, R_0 [9]. If infection started in the under 20 age group, there was a noticeable bias in the outbreak size distribution, with large outbreaks in under 20 year-olds more likely than large outbreaks in the over 20s
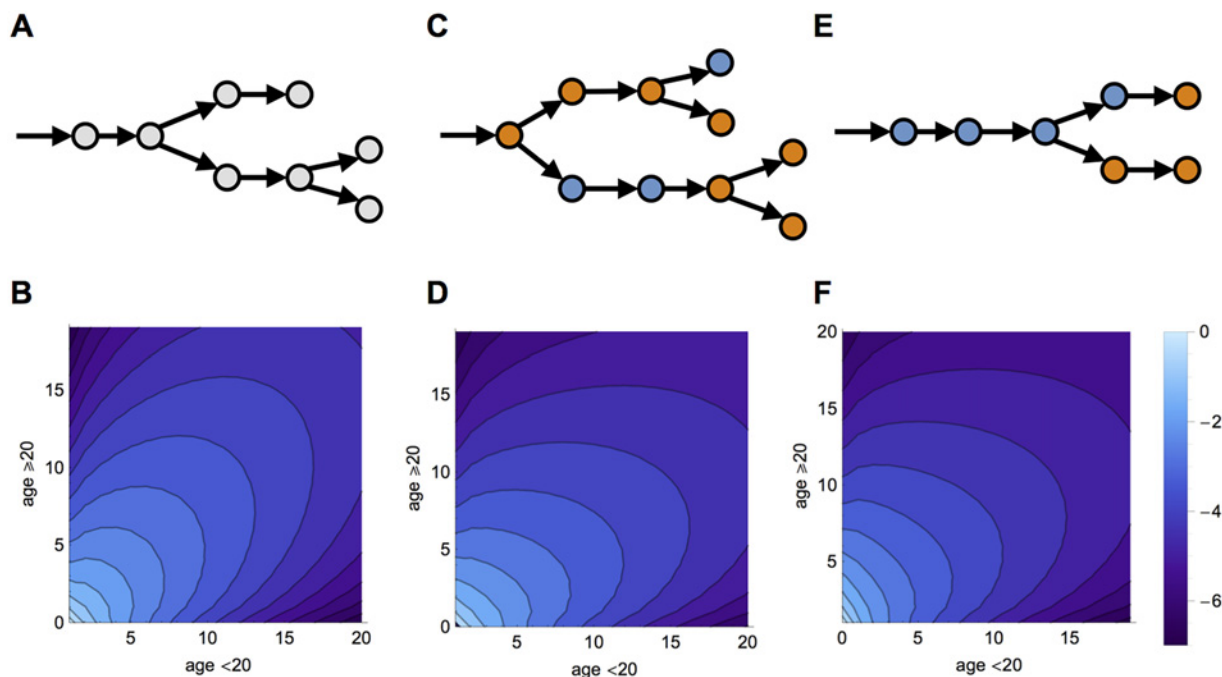
Fig 1. Transmission chains and joint outbreak distributions. (A) Example transmission chain when the population mixes homogeneously. (B) Joint probability an introduction will produce a transmission chain of a given size in each of the two age groups (on $\log_{10}$ scale) when the outbreak starts in the under 20 age group. (C) Transmission chain when mixing is age-dependent and infection starts in the under 20 age group. (D) Joint outbreak size distribution when model incorporates social contact data from Great Britain and infection introduced into the under 20 age group.(E) Transmission chain when infection starts in oldest age group. (F) Joint size distribution when infection starts in the over 20 age group. We assume $R_0 = 0.6$.

doi:10.1371/journal.pcbi.1004154.g001

([Fig. 1D](#)). When the infection started in the over 20 age group ([Fig. 1E](#)), the offspring distribution shifted, and the probability of large outbreaks in the under 20 age group decreased ([Fig. 1F](#)).

Identifying anomalously large outbreaks

We used the outbreak size distribution to identify what constitutes an anomalously large outbreak for a particular R_0. We defined this as an outbreak size that has a less than 10^{-3} probability of occurring in our model. When the infection was introduced into the under 20 age group, there was an asymmetry in the threshold for an unusually large outbreak in the UK ([Fig. 2A](#)). If $R_0 = 0.7$, a chain of at least 8 cases was not unusual if some of the secondary cases are children, yet it is if the secondary cases are all adults. The conditions for an anomalously large outbreak shifted when infection started in the eldest group ([Fig. 2B](#)). In some cases the thresholds curved inwards. In [Fig. 2A](#), when $R_0 = 0.7$ an outbreak of size 7 was anomalously large if all secondary cases were in the youngest group, but an outbreak of size 10 was not unusual if between 2–8 secondary cases were in the eldest group. As the infection was introduced in the youngest group, this suggested that chains of transmission were more likely to persist if they crossed into the eldest age group. The threshold also curved inwards when the infection started in the eldest group ([Fig. 2B](#)). An outbreak of size 5 was unusual if all the secondary cases were in the youngest group, but an outbreak of size 8 was not anomalous if there were 3 cases in the eldest group. This implies that having a single case in the introductory age group and several in the other group was unlikely when $R_0 = 0.7$. As suggested by the next generation matrix ([S1A Fig](#).), the primary case would generally create additional cases within the same group rather than infect only individuals in the other group.

Estimating transmissibility and pre-existing immunity

Using age stratified-data, we found that is was possible to distinguish between inherent pathogen transmissibility and pre-existing host immunity. We simulated outbreaks using a multitype branching process with two groups, then used the outbreak size distribution to infer R_0 and relative immunity in older individuals. We assumed that the under 20 age group was fully
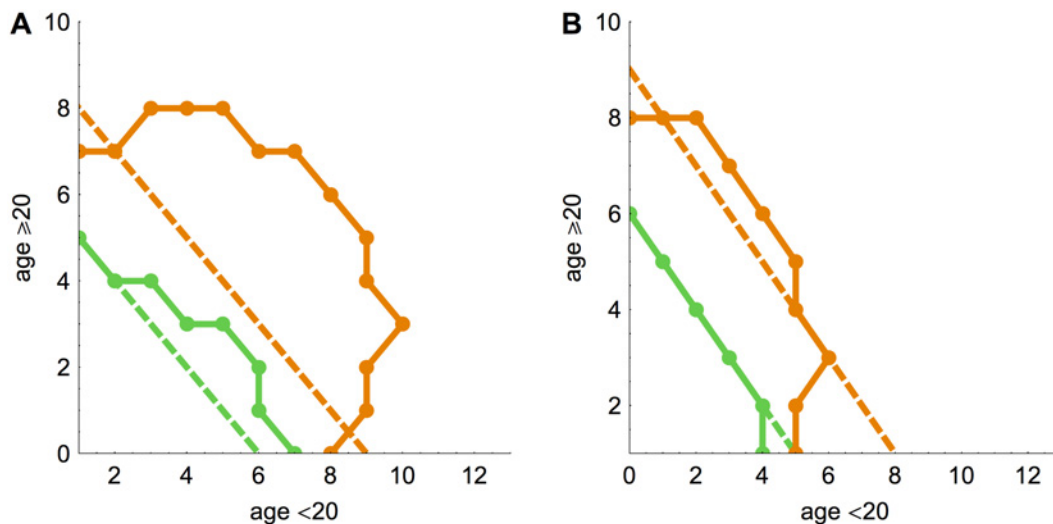

Fig 2. Threshold for anomalously large outbreak sizes. (A) Primary case is in the under 20 age group. Points show joint outbreak sizes that have less than 10^{-3} probability of occurring. Green points, $R_0 = 0.3$; orange, $R_0 = 0.7$. (B) Primary case is in the over 20 age group.

doi:10.1371/journal.pcbi.1004154.g002

susceptible to infection, and the relative susceptibility of the over 20 age group, denoted S, could vary. Each outbreak was seeded randomly in the susceptible population. In the UK the under 20 age group make up 24% of the total population, so in the absence of immunity, the probability of the outbreak starting in this group was 0.24.

To test our inference framework, we simulated four different scenarios. First, we examined two infections with the same $R_0 = 0.2$, but different levels of immunity in the over 20 age group. In one scenario, only 20% of hosts over age 20 were susceptible to infection (i.e. $S = 0.2$); in the other, the population was fully susceptible ($S = 1$). We simulated 50 spillover events, and found the maximum likelihood estimate of R_0 and S. We repeated this process for 1000 sets of outbreaks, obtaining reliable estimates of both R_0 and S (Figs. 3A-B). Next, we considered the same two susceptibility values, but for an infection with $R_0 = 0.7$. The model was again able to distinguish between the different scenarios (Figs. 3C-D). The structure of the reproduction matrix (Equation 2) means that R_0 and S should always be identifiable in the model, given enough
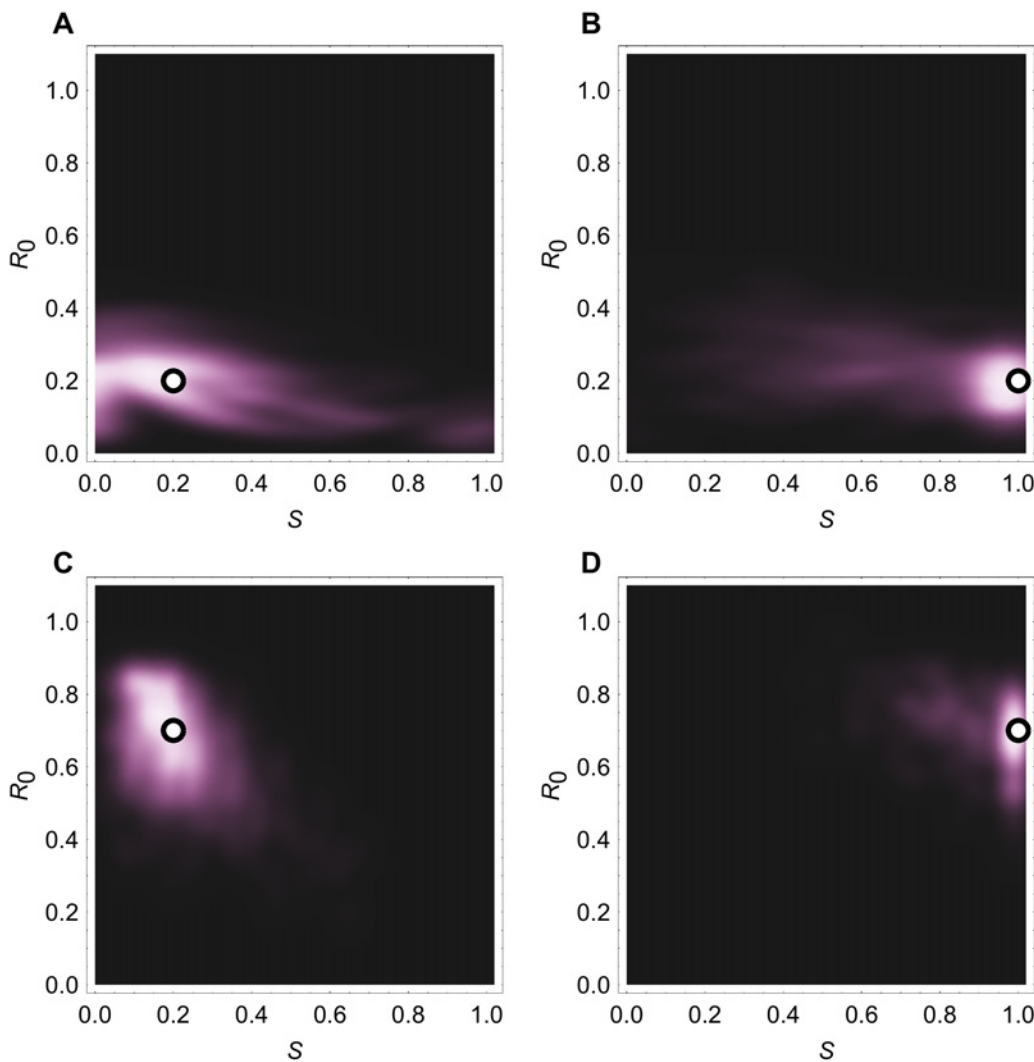

Fig 3. Estimates of R_0 and relative susceptibility of over 20 age group, S, from simulated data. We simulated 1000 sets of 50 outbreaks, and found the maximum likelihood estimates (MLEs) for parameters for each set. White dots show true parameter values; heat map shows distribution of the 1000 MLEs.

doi:10.1371/journal.pcbi.1004154.g003

data, because R_0 scales the entire matrix, whereas S only scales the transmission rate to the older age group.

We used our estimates of R_0 and relative immunity in the over 20 age group to calculate the effective reproduction number. We compared these values with estimates from an inference framework based on a single-type branching process [15, 16, 17, 18]. In all four scenarios, our estimates for R are less biased in the age-structured model (Table 1). However, the relative sum-squared error is smaller in the single-type model when R_0 is small. This is because accurate inference across the two age groups requires sampling from the tail of the joint outbreak size distribution, which is achieved either when R_0 is larger (Table 1), or when more outbreak data are available. When inference is performed using data from a larger number of outbreaks, the relative error for the age-structured model is smaller than for the non-stratified framework (S2 Fig.).

Regardless of the degree of transmissibility or immunity, we systemically underestimate R in the single-type model (Table 1). This bias is the result of our assumption that introductions occur randomly across the susceptible population, and illustrates an important caveat to inference of R from the mean outbreak size in a single-type branching process model. If the proportion of cases that are introduced to each age group is equal to the dominant eigenvector of the reproduction matrix, it is possible to obtain unbiased estimates for R using only the mean outbreak size (see Text S1). However, if the true proportion of introductions in the under 20 group is less than number of introductions implied by dominant eigenvector, we will underestimate R in a single-type model (S3 Fig.). Conversely, if the true proportion of introductions is larger, we overestimate R. In our model of transmission chains in Great Britain, we assumed a child-dominated social contact matrix but relatively flat population structure. In the absence of immunity, the probability the infection starts in the under 20 age group was therefore 0.24. However, the relevant component of the dominant eigenvector of the reproduction matrix is 0.68. If the probability of introduction is less than this—as it is in our model—the homogeneous mixing assumption will lead to an underestimate of R (S3 Fig.). The age structured model avoids dependency on age-specific exposure risk by accounting for which age group the infection started in when performing inference (Equation 13). If there were a disproportionate number of introductions in a particular age group, the structure of the likelihood function means that it would not bias our estimate for R.

We also tested whether our inference approach, which assumed social contact data reflects age-specific transmission, was sensitive to misspecification of the 'true' transmission process. We simulated data using different assumptions about age-specific infection rates but left the

Table 1. Accuracy of estimation of effective reproduction number, R.

R_0	S	R	Model	R bias	Relative R error
0.2	0.2	0.16	Single-type	-0.026	0.36
			Age-structured	-0.002	0.38
0.2	1	0.2	Single-type	-0.030	0.32
			Age-structured	0.008	0.34
0.7	0.2	0.56	Single-type	-0.066	0.21
			Age-structured	-0.021	0.17
0.7	1	0.7	Single-type	-0.070	0.17
			Age-structured	-0.014	0.12

We simulated 1000 sets of 50 spillover events, and found the maximum likelihood estimates of the reproduction number, $\hat{R}$, for each set.

doi:10.1371/journal.pcbi.1004154.t001

inference model unchanged. First, we simulated outbreak data using a multi-type branching process with 15 age groups. As in the inference model, transmission between different groups depended on reported physical contacts from the POLYMOD survey in Great Britain. Although the inference model only used two age groups, it correctly identified the four different combinations of transmissibility and susceptibility (S4 Fig.). Next, we simulated data using two ages groups, but with transmission based on the average number of reported physical contacts across 8 European countries in the POLYMOD study (S1B Fig.). The relative error in R was generally slightly larger (S1 Table), but we were still able to obtain accurate scenario estimates (S5 Fig.). When we considered a generic child-dominated next generation matrix (S1C Fig.), our estimates for S were more variable, but we were still able to distinguish between pathogen transmissibility and pre-existing immunity (S6 Fig.). Finally, we considered a transmission matrix in which adults were dominant (S1D Fig.). As expected in such a heavily mis-specified model, we were not able to accuracy estimate S and R_0 (S7 Fig.).

Application to real outbreaks

Using our age-stratified framework, we characterized the transmission potential of four infections (Table 2): influenza A(H5N1); influenza A(H7N9); Monkeypox; and MERS-CoV. As we could not be certain that the under 20 age group was fully susceptible, we did not infer the basic reproduction number, R_0. Instead, we defined ρ to be the effective reproduction number when both groups were equally susceptible (i.e. $S = 1$). If in reality the under 20 age group had no immunity to the infection then $\rho = R_0$. For our analysis of MERS and monkeypox outbreak data, we used the average reported physical contacts from POLYMOD across 8 European countries (S1B Fig.). For H5N1 and H7N9, we used physical contact data from Southern China (S1E Fig.).

We measured transmission potential by jointly inferring ρ and S for each of the four infections. Our maximum likelihood estimates suggest that the over 20 age group had substantial pre-existing immunity against monkeypox and H5N1, and no immunity against H7N9 or MERS-CoV (Fig. 4). These estimates agree with values derived from detailed studies of vaccination and infection history (Table 3). We could not perform such a comparison for MERS-CoV, however, as we could find no studies reporting measurements of population-level immunity for humans.

Our estimate of S for monkeypox exhibited considerable uncertainty: the 95% confidence interval spanned 0.02–1. This was likely the result of the small number of clusters we analysed. To examine whether a larger number of clusters might improve our model estimates, we performed a simulation study using an infection with limited transmissibility in a population with pre-existing immunity (i.e. a similar scenario to monkeypox transmission). We simulated 50 spillover events, with $R_0 = 0.25$ and $S = 0.5$, then attempted to infer the parameters from the age-stratified outbreak size data. We found that the 95% confidence interval of the joint

Table 2. Details of outbreaks analysed.

Infection	Location	Period	Cases	Clusters	Source
Monkeypox	Central Africa	24/08/70–30/10/79	47	41	[5]
Influenza A(H5N1)	Indonesia	21/07/05–22/11/13	192	176	WHO, [43]
Influenza A(H7N9)	Shanghai & Jiangsu	19/02/13–30/04/13	59	55	[44]
MERS	Globally	20/09/12–01/08/13	110	42	[4]

Clusters denote sets of human cases that have a known epidemiological link; single cases are defined as clusters of size one.

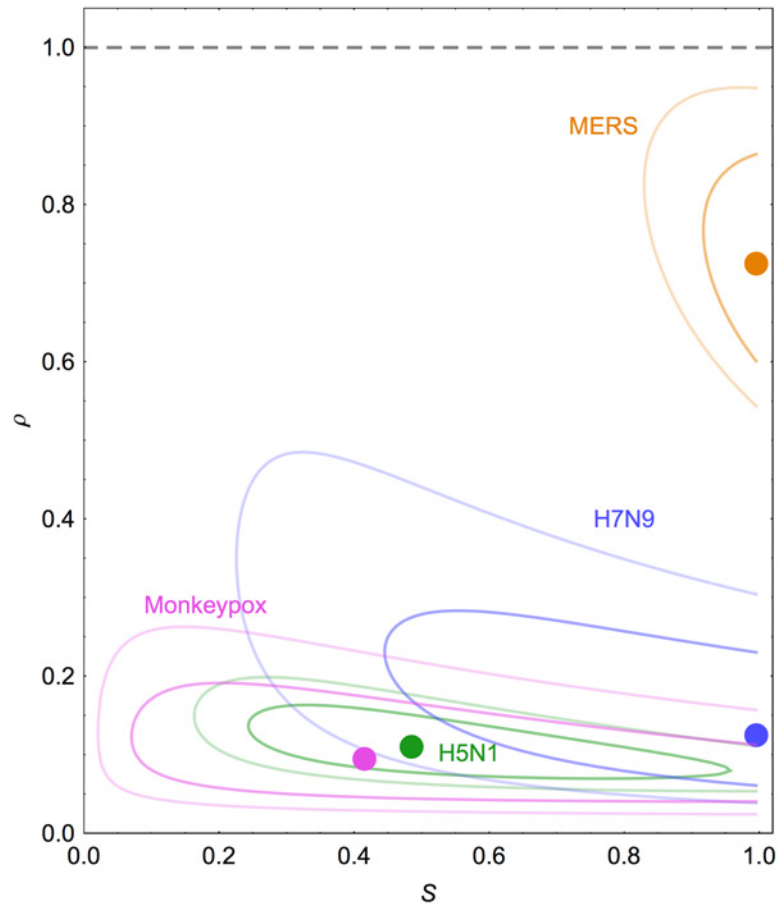doi:10.1371/journal.pcbi.1004154.t002

Fig 4. Characterization of transmission potential from observed outbreak size distributions. Each point shows joint maximum likelihood estimate of the effective reproduction number if both age groups were equally susceptible, ρ, and the relative susceptibility of over 20s, S. Dark line indicates 80% confidence interval (CI); light line is 95% CI. Blue, influenza A(H7N9); green, influenza A(H5N1); pink, monkeypox; orange, MERS.

doi:10.1371/journal.pcbi.1004154.g004

Table 3. Comparison of our parameter estimates and previously published estimates of *R*, from studies using single-type models, and *S*, from studies of vaccination history [5] and seroprevalence [28, 29] and models of immune acquisition [27].

Infection	R estimate (95% CI)	R in literature	S estimate	S in literature
Monkeypox	0.08 (0.02–0.22)	0.09 [5]	0.42 (0.02–1)	0.4–0.7* [5]
H5N1	0.10 (0.05–0.17)	0.14 (0.004–0.39) [7]	0.48 (0.15–1)	0.36 (0.30–0.44)* [27]
H7N9	0.12 (0.04–0.42)	0.10 (0.01–0.49) [45]	0.99 (0.22–1)	≈1 [28, 29]
MERS	0.73 (0.54–0.96)	0.63 (0.49–0.85) [3, 4]	0.99 (0.82–1)	–

*See Text S1 for details of calculation.

doi:10.1371/journal.pcbi.1004154.t003

distribution of R_0 and S was very broad (S8 Fig.). However, when we simulated 150 or 250 spill-over events instead, the uncertainty in our estimates shrank, and we were able to obtain more precise parameter estimates (S8 Fig.).

Using our model, we also estimated R for each set of real outbreaks. Our estimates were similar to previously published estimates that assumed a single-type population. However, the confidence intervals for our estimates were generally smaller (Table 3). Influenza A(H7N9) had an effective reproduction number of 0.08 (95% CI 0.02–0.23), influenza A(H5N1) had $R = 0.10$ (0.05–0.18) and monkeypox $R = 0.08$ (0.02–0.22). Our estimate of R for MERS-CoV was 0.73 (0.54–0.96), whereas in a single-type branching process model $R = 0.63$ (0.49–0.85). The discrepancy was caused by the age distribution of the largest outbreak clusters. One cluster of 26 infections consisted entirely of over 20s: if transmission was indeed driven by social mixing patterns, such an outbreak would require a large R to persist in only one group.

During real-time analysis of an outbreak, there may be additional infections yet to be reported. It is particularly important to account for such censoring when infections are near the $R = 1$ boundary [17]. To test the robustness of our estimates for MERS-CoV when outbreak size data were censored, we extended our inference framework to account for incomplete outbreaks (methods in Text S1). When censoring was included, our estimate for R increased slightly to 0.77 (0.57–1.03), but our maximum likelihood estimate for S remained the same.

Discussion

Obtaining accurate estimates of transmission potential is crucial for effective surveillance and control of infectious diseases. However, for emerging infections estimates often have to be made using case data from a limited number of small outbreaks. Using a multi-type branching process, we developed an inference framework to make better use of age-structured outbreak size data.

Our results show that when disease transmission is driven by social contacts between different age groups, knowledge of the age distribution of cases makes it possible to separate the effective reproduction number into two components: inherent pathogen transmissibility, and pre-existing immunity in older age groups. Based on observed outbreak size distributions, we estimated that individuals over age 20 had susceptibility to monkeypox reduced by a factor 0.4 compared with younger hosts. This value agrees well with published estimates of population susceptibility (Table 3), with cross-immunity coming from the smallpox vaccination campaigns that ended in the two decades preceding the outbreaks [5]. We also found evidence of pre-existing immunity to influenza A(H5N1) in older individuals; it has previously been suggested that such immunity could result either from prior exposure to H5N1, or from cross-immunity from previous infection with influenza A(H1N1) [27]. In contrast, we estimated that both age groups had similar levels of susceptibility to MERS and influenza A(H7N9). Given that immunity from vaccination and natural infection tends to increase with age, this suggests that there was little pre-existing immunity to these pathogens. While serological studies have found no evidence of pre-existing immunity to H7N9 virus in these locations [28, 29], serological analysis remains challenging for novel coronaviruses such as MERS-CoV [30]. The approach we describe can therefore provide crucial information about the degree of population susceptibility before serological surveys are available.

We also used our model to identify thresholds for anomalously large outbreaks. In a single-type branching process framework, the threshold is a single number: the total size of the outbreak [31, 16]. In an age-structured model, however, the threshold depends on outbreak size in each age group. The age breakdown of cases can therefore provide additional information about what constitutes an unusual outbreak which would not be available with only overall

outbreak sizes. Moreover, the shape of the thresholds in Fig. 2 suggest that the infection must pass between age groups to persist. Such dynamics could be important in understanding how pathogens adapt to a new host or invade a new population, and could be explored further in future using the models we have described here.

We made several assumptions in our model. First, we assumed that secondary cases are drawn from a geometric distribution with mean R (or R_{ij} in the two group model). This is akin to assuming that recovery times are exponentially distributed in the standard SIR model. Other studies have assumed that the offspring distribution for secondary cases follows a negative binomial distribution, and have suggested that an increased level of over-dispersion is often appropriate when modelling disease emergence [16, 14]. However, some of this over-dispersion is captured implicitly our model as a result of the variation that comes from including social contact structure. Given appropriate data, it would be interesting to see whether individual variation in transmission can be explained by social behaviour rather than processes such as virus shedding. This would have implications for how the over-dispersion parameter should be interpreted in an age-structured framework.

We also assumed that transmission events are independent, and did not consider depletion of susceptible hosts during an outbreak. This simplification is reasonable for infections with a small effective reproduction number, but depletion of susceptibles would need to be accounted for if R were close to 1 [16]. In addition, we assumed that transmission potential between age groups was captured entirely by social contacts. Because we used simulated data to infer parameters, and hence had knowledge of the true model, we were also assuming that these contacts were reported accurately. We tested the accuracy of parameter estimation when the transmission process was mis-specified, and found that it was still possible to distinguish between different scenarios as long as transmission matrices in both the simulation and inference models were dominated by intense mixing between children. This is a reasonable assumption, as it has been suggested that such mixing patterns drive observed outbreaks of respiratory infections [21, 25, 32].

Although published contact matrix data were not available for Central Africa, where monkeypox cases were reported, preliminary results from social contact survey in Uganda suggest that age mixing patterns are qualitatively similar to those found in the POLYMOD study, with a clear pattern of assortative mixing between different age groups, and children reporting a larger number of contacts relative to adults (Olivier Le Polain de Waroux, personal communication).

Our work extends existing techniques for inferring epidemiological parameters from the distribution outbreak sizes. By accounting for the age structure of a population, we show that it is possible to obtain unbiased estimates of the reproduction number, and distinguish between pathogen transmissibility and immunity from outbreak size data. During an outbreak, cluster data may be difficult to obtain; cases are typically reported as aggregated totals by health ministries and WHO [33]. Our results illustrate the value of making higher resolution outbreak data available, with cluster information and covariates such as age reported along overall case numbers.

There are situations in which it could be necessary to distinguish inherent pathogen transmission potential from immunity. For example, if a vaccination campaign that protects against an infection is to change, or be discontinued, it would be important to understand how the pathogen could transmit in a fully susceptible population. This question motivated early studies of monkeypox transmission [34]. However, in studies of monkeypox outbreaks it was relatively straightforward to identify a case's smallpox vaccination history, because the smallpox vaccine—which provided cross-immunity to monkeypox—left a distinctive scar. The same might not to be true for other vaccines.

Our methods are not limited to age structure, and could be used to examine a variety of population stratifications. Depending on the pathogen, transmission rates may also depend on factors such as profession or setting (for example, hospital versus community transmission). With appropriately stratified outbreak data, it would be possible to infer relative immunity and transmissibility in range of different groups. While spillover infections such as avian influenza and MERS-CoV are a natural application for our approach, population structure could also influence the dynamics of transmission chains following introduction via other routes. For example, novel pathogen strains could emerge via resistance-conferring mutations [35] or adaptation to a human host [36], or be introduced to a population through air travel [37]. By collecting secondary information such as the age distribution of cases, and combining these data with models such as the one outlined here, it should be possible to develop a better understanding of stuttering chains of infection and their transmission potential. During an outbreak, our framework would also be able to generate estimates of epidemiological parameters from a commonly available data source, and hence characterize transmission risk before serological surveys and other detailed analyses are available.

Methods

Data

Contact data came from the POLYMOD study, a diary-based survey conducted in Europe [20], and a study of social mixing patterns in Southern China [38]. In both studies, participants reported the age of their contacts on a specified day, defined as either a face-to-face conversation in the physical presence of another person, or physical skin-to-skin contact. In our simulation study, we used data on reported physical contacts from the POLYMOD survey in Great Britain (S1A Fig.) to define the level of transmission between different age groups, as there is evidence that this type of contact is better proxy for respiratory pathogen transmission than total contacts [25, 32]. Similar qualitative mixing patterns can be found in other European countries (S1B Fig.) and Southern China (S1E Fig.), as well as Southeast Asian countries such as Vietnam [39] and Hong Kong [25]. Outbreak size distributions for different infections were calculated from reported cases (Table 2). In the influenza A(H5N1) data, it was not always clear whether an outbreak cluster was seeded by a single primary case—with all other infections secondary—or multiple co-primary cases. We made the conservative assumption that each cluster had only one primary case: our estimate for R can therefore be considered to be an upper bound on potential transmissibility given available outbreak size data.

Next generation matrix

We used the next generation matrix to describe the average number of secondary cases in a population with two age groups. To model age-dependent infection, we defined m_{ij} to be the mean number of contacts with individuals in age group i reported by participants in age group j, and λ to be the maximal eigenvalue of the matrix **M** with entries m_{ij}. Defining S to be the relative susceptibly of group 2 compared to group 1, the average number of infections to group i from group j was therefore given by [40]:

$$R_{ij} = \begin{cases} qm_{ij}/\lambda & \text{if } i = 1 \\ qSm_{ij}/\lambda & \text{if } i = 2 \end{cases} \tag{1}$$

where q is a scaling factor depending on inherent pathogen transmissibility (i.e. R_0). We defined the next generation matrix, $\mathbf{R}$, to be the matrix with entries R_{ij},

$$\mathbf{R} = \begin{pmatrix} R_{11} & R_{21} \\ R_{12} & R_{22} \end{pmatrix} . \tag{2}$$

The effective reproduction number of the infection, R, was equal to the dominant eigenvalue of this matrix. If the population was fully susceptible, then R was equal to the basic reproduction number, R_0. If $S = 1$, but we did not know whether the population as a whole was fully susceptible, then we defined the dominant eigenvalue to be ρ.

Offspring distribution

We used a multi-type branching process to model secondary infections (see Text S1 for details). Given two different types of individuals, the generating function for the offspring distribution of individual i was

$$h_i(s_1, s_2) = \sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} p_{s_1, s_2} s_1^{j_1} s_2^{j_2} \tag{3}$$

where p_{s_1, s_2} was the probability that an infectious individual of type i generated s_1 secondary cases of type 1 and s_2 cases of type 2. We assumed that stochasticity in transmission was represented by a Poisson process, and that the individual offspring distribution followed a negative binomial distribution [14]:

$$h_i(s_1, s_2) = \left(1 + \frac{R_{1i}}{k}(1 - s_1)\right)^{-k} \left(1 + \frac{R_{2i}}{k}(1 - s_2)\right)^{-k} . \tag{4}$$

It was possible to separate this probability generating function into two components,

$$h_i(s_1, s_2) = g_{1i}(s_1) g_{2i}(s_2) . \tag{5}$$

Extending approaches used for a single-type population [16], we could specify the probability that a certain number of cases of type i are generated by infectives of type j (see Text S1 for details):

$$\mathbb{P}(z \text{ cases of type } i \text{ generated by } n \text{ cases of type } j) = T_{ij}^n(z) \tag{6}$$

$$= \frac{1}{z!} \frac{d^z}{ds^z} [g_{ij}(s)]^n \Big|_{s=0} . \tag{7}$$

Inserting the relevant part of Equation 4 into Equation 7, we obtained

$$T_{ij}^n(z) = \frac{\prod_{w=0}^{z-1}(kn + w)}{z!} \left(\frac{R_{ij}}{k}\right)^z \left(1 + \frac{R_{ij}}{k}\right)^{-z-kn} . \tag{8}$$

Note that in this paper we set $k = 1$. This was equivalent to assuming that recovery times were exponentially distributed, as in the standard SIR model.

Outbreak size distribution

We used the offspring distribution to calculate the probability that an outbreak results in the following outcome: n total cases in group 1; m total cases in group 2; a_{12} infections in group 1 caused by infective hosts in group 2; and a_{21} infections in group 2 caused by infective hosts in

group 1. There were two situations to consider. If $m = 0$, then $a_{12} = a_{21} = 0$ and hence [23],

$$P(n, m) = \frac{T_{11}^n(n-1) T_{21}^n(0)}{n} \quad .$$

(9)

If $m > 0$, we had [24],

$$P(n, m, a_{21}, a_{12}) = \frac{a_{21} T_{11}^n(n - a_{12} - 1) T_{21}^n(a_{21}) T_{12}^m(a_{12}) T_{22}^m(m - a_{21})}{nm} \quad .$$

(10)

Finally, we used Equations 9–10 to calculate $r_{n,m}^1$, the probability the infection will cause an outbreak of size n in group 1 and m in group 2, given that the initial case was in group 1:

$$r_{n,m}^1 = \sum_{a_{12}=0}^{A_2} \sum_{a_{21}=0}^{m} P(n, m, a_{21}, a_{12})$$

(11)

where

$$A_2 = \begin{cases} 0 & \text{if } m = 0 \\ n - 1 & \text{else} \end{cases} \quad .$$

(12)

By symmetry, we can obtain an analogous expression for $r_{n,m}^2$.

Inference

If $N_{n,m}^i$ was the number of chains that start in group i and resulted in n cases in group 1 and m cases in group 2, then by Equation 11 the likelihood of parameter set θ given data X was:

$$L(\theta \mid X) = \prod_{m=1}^{\infty} \prod_{n=1}^{\infty} (r_{n,m}^1)^{N_{n,m}^1} (r_{n,m}^2)^{N_{n,m}^2} \quad .$$

(13)

When only the total number of cases in a cluster was known, and not the age distribution, we instead inferred the reproduction number from the overall outbreak size distribution [16]. If N_n was the number of chains of size n, and r_n was the probability a transmission chain has size n, the likelihood function was:

$$L(\theta \mid X) = \prod_{j=1}^{\infty} r_n^{N_n} \quad .$$

(14)

We obtained maximum likelihood estimates for $\theta = \{R_0, S\}$ by calculating the two-dimensional likelihood surface and using a simple grid-search algorithm to find the maximum point. For a higher dimensional model, it might be necessary to use an alternative technique, such as Markov chain Monte Carlo [41], to ensure robust and efficient parameter estimation. Confidence intervals were calculated using profile likelihoods: for each value of R_0, we found the maximum likelihood across all possible values of S; the 95% confidence interval was equivalent to the region of parameter space that was within 1.92 log-likelihood points of the maximum-likelihood estimate for both parameters [42].

Performance metrics

It was not possible to obtain a tractable expression for the maximum likelihood (ML) estimates of ρ and S, and hence R, using Equation 13. Instead we calculated the ML estimate of the reproduction number, $\hat{R}$, using the numerically estimated maximum likelihood values for ρ and S. We used two metrics to assess the accuracy of $\hat{R}$: the estimator bias and relative error [15].

Having generated M sets of outbreak data using the same R, and found $\hat{R}_i$ for each set i, the estimator bias was

$$\delta_1 = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \hat{R}_i - R \, , \tag{15}$$

and the root mean square relative error was given by:

$$\delta_2 = \sqrt{\lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \left(\frac{\hat{R}_i - R}{R} \right)^2} \, . \tag{16}$$

Mean outbreak size in two group model

Let $\boldsymbol{\mu}$ denote the mean outbreak size matrix. If we denote entries of $\boldsymbol{\mu}$ by μ_{ij}, then $\Sigma_j \, \mu_{ij}$ is the mean outbreak size in group i. If the eigenvalues of the next generation matrix $\mathbf{R}$, denoted λ_i, are such that $|\lambda_i| < 1$ for all i, we have

$$\boldsymbol{\mu} = \mathbf{IA} + \mathbf{RA} + \mathbf{R}^2\mathbf{A} + \ldots = \sum_{k=0}^{\infty} \mathbf{R}^k \mathbf{A} = (\mathbf{I} - \mathbf{R})^{-1}\mathbf{A} \tag{17}$$

where

$$\mathbf{A} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \tag{18}$$

and σ_i is the probability the primary infection was in group i.

Supporting Information

S1 Fig. Contact matrix data used in model. (A) Reported physical contacts in Great Britain in POLYMOD study [20], (B) Average across 8 European countries [20], (C) Example child-dominated matrix, (D) Example adult-dominated matrix, (E) Reported physical contacts in Southern China [38].
(TIFF)

S2 Fig. Simulation results for error as number of chains increases. (A) $R_0 = 0.2$ and $S = 0.2$. Blue line, relative error in maximum likelihood estimate for R in single-type model; red line, error in estimate for R in age-structured model. (B) $R_0 = 0.2$ and $S = 1$.
(TIFF)

S3 Fig. Inferred value of R using overall mean outbreak size (Equation 17). Blue line, population fully susceptible ($S = 1$); green line, over 20 age group have susceptibility reduced by half relative to under 20 group ($S = 0.5$). If the probability that the infection is introduced into group 1 (i.e. under 20 age group)
(TIFF)

S4 Fig. Estimates of R_0 and relative susceptibility, S, when simulation model is a multi-type branching process with 15 age groups. We simulated 1000 sets of 50 outbreaks, and found the maximum likelihood estimates (MLEs) for parameters for each set. White dots show true parameter values; heat map shows distribution of the 1000 MLEs.
(TIFF)

S5 Fig. Estimates of R_0 and relative susceptibility, S, when inference model assume GB contact patterns and simulation model uses average mixing patterns across 8 European countries ([S1B Fig](#).). We simulated 1000 sets of 50 outbreaks, and found the maximum likelihood estimates (MLEs) for parameters for each set. White dots show true parameter values; heat map shows distribution of the 1000 MLEs.
(TIFF)

S6 Fig. Estimates of R_0 and relative susceptibility, S, when inference model assume GB contact patterns and simulation model uses generic child-dominated next generation matrix ([S1C Fig](#).). We simulated 1000 sets of 50 outbreaks, and found the maximum likelihood estimates (MLEs) for parameters for each set. White dots show true parameter values; heat map shows distribution of the 1000 MLEs.
(TIFF)

S7 Fig. Estimates of R_0 and relative susceptibility, S, when inference model assume GB contact patterns and simulation model uses generic adult-dominated next generation matrix ([S1D Fig](#).). We simulated 1000 sets of 50 outbreaks, and found the maximum likelihood estimates (MLEs) for parameters for each set. White dots show true parameter values; heat map shows distribution of the 1000 MLEs.
(TIFF)

S8 Fig. Estimates of R_0 and relative susceptibility, S, as number of spillover events increased. In simulations, $R_0 = 0.25$ and $S = 0.5$. Age-specific contact patterns were based on reported physical contacts in Great Britain in POLYMOD study [20].
(TIFF)

S1 Table. Accuracy of R estimation when inference matrix is mis-specified (Matrix in [S1B Fig](#).).
(PDF)

S1 Code. Simulation and inference code. Simulation model generates stochastic multi type outbreaks from a two-class mixing matrix. The inference model generates maximum likelihood estimates of R_0 and S from outbreak size data.
(R)

Acknowledgments

We would like to thank Paul Fine, Theo Kypraios and Jamie Lloyd-Smith for useful discussions.

Author Contributions

Conceived and designed the experiments: AJK WJE. Performed the experiments: AJK. Analyzed the data: AJK WJE. Wrote the paper: AJK WJE.

References

1. Anderson RM, Fraser C, Ghani AC, Donnelly CA, Riley S, Ferguson NM, et al. Epidemiology, trans mission dynamics and control of SARS: the 2002–2003 epidemic. Philos Trans R Soc Lond B Biol Sci. 2004; 359 (1447):1091–105. doi: 10.1098/rstb.2004.1490 PMID: 15306395

2. Worobey M, Han GZ, Rambaut A. Genesis and pathogenesis of the 1918 pandemic H1N1 influenza A virus. Proc Natl Acad Sci USA. 2014; 111(22):8107–12. doi: 10.1073/pnas.1324197111 PMID: 24778238

3. Breban R, Riou J, Fontanet A. Interhuman transmissibility of Middle East respiratory syndrome coro navirus: estimation of pandemic risk. The Lancet. 2013; 382 (9893):694–699. doi: 10.1016/S0140-6736(13)61492-0

4. Cauchemez S, Fraser C, Van Kerkhove MD, Donnelly CA, Riley S, Rambaut A, et al. Middle East respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility. The Lancet Infectious Diseases. 2014; 14(1):50–56. doi: 10.1016/S1473-3099(13)70304-9 PMID: 24239323

5. Breman J, Kalisa-Ruti M, Zanotto E, Gromyko A, Arita I. Human monkeypox, 1970–79. Bulletin of the World Health Organization. 1980; 58(2):165. PMID: 6249508

6. Lloyd-Smith JO, George D, Pepin KM, Pitzer VE, Pulliam JRC, Dobson AP, et al. Epidemic dynamics at the human-animal interface. Science. 2009; 326(5958):1362–7. doi: 10.1126/science.1177345 PMID: 19965751

7. Aditama TY, Samaan G, Kusriastuti R, et al. Avian influenza H5N1 transmission in households, Indonesia. PLoS One. 2012; 7(1):e29971. doi: 10.1371/journal.pone.0029971 PMID: 22238686

8. Li Q, Zhou L, Zhou M, Chen Z, Li F, Wu H, et al. Epidemiology of human infections with avian influenza A(H7N9) virus in China. N Engl J Med. 2014; 370(6):520–32. doi: 10.1056/NEJMoa1304617 PMID: 23614499

9. Anderson RM, May RM. Directly transmitted infectious diseases: control by vaccination. Science.1982; 215 (4536):1053–1060. doi: 10.1126/science.7063839 PMID: 7063839

10. Rimoin AW, Mulembakani PM, Johnston SC, Lloyd Smith JO, Kisalu NK, Kinkela TL, et al. Major increase in human monkeypox incidence 30 years after smallpox vaccination campaigns cease in the Democratic Republic of Congo. Proc Natl Acad Sci U S A. 2010 Sep; 107(37):16262–7. doi: 10.1073/pnas.1005769107 PMID: 20805472

11. Miller E, Hoschler K, Hardelid P, Stanford E, Andrews N, Zambon M. Incidence of 2009 pandemic influenza A H1N1 infection in England: a cross-sectional serological study. The Lancet. 2010; 375(1100–1108:).

12. Grenfell B, Anderson R. The estimation of age-related rates of infection from case notifications and serological data. Epidemiology and Infection. 1985; 95(2):419–436.

13. De Serres G, Gay NJ, Farrington C. Epidemiology of transmissible diseases after elimination. American Journal of Epidemiology. 2000; 151(11):1039–1048. doi: 10.1093/oxfordjournals.aje.a010145

14. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005; 438 (7066):355–9. doi: 10.1038/nature04153 PMID: 16292310

15. Blumberg S, Lloyd-Smith J. Comparing methods for estimating R_0 from the size distribution of subcritical transmission chains. Epidemics. 2013; 5(3):131–45. doi: 10.1016/j.epidem.2013.05.002 PMID: 24021520

16. Blumberg S, Lloyd-Smith JO. Inference of R_0 and transmission heterogeneity from the size distribution of stuttering chains. PLoS Comput Biol. 2013; 9(5):e1002993. doi: 10.1371/journal.pcbi.1002993 PMID: 23658504

17. Farrington C, Kanaan M, Gay N. Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics. 2003; 4(2):279. doi: 10.1093/biostatistics/4.2.279 PMID: 12925522

18. Nishiura H, Yan P, Sleeman CK, Mode CJ. Estimating the transmission potential of supercritical processes based on the final size distribution of minor outbreaks. J Theor Biol. 2012; 294:48–55. doi: 10.1016/j.jtbi.2011.10.039 PMID: 22079419

19. Edmunds WJ, O'Callaghan CJ, Nokes DJ. Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. Proc Biol Sci. 1997; 264 (1384):949–57. doi: 10.1098/rspb.1997.0131 PMID: 9263464

20. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 2008; 5(3):e74. doi: 10.1371/journal.pmed.0050074 PMID: 18366252

21. Wallinga J, Teunis P, Kretzschmar M. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. American Journal of Epidemiology. 2006; 164(10):936. doi: 10.1093/aje/kwj317 PMID: 16968863

22. Flasche S, Hens N, Boëlle PY, Mossong J, et al. Different transmission patterns in the early stages of the influenza A(H1N1)v pandemic: a comparative analysis of 12 European countries. Epidemics.2011; 3(2):125–33. doi: 10.1016/j.epidem.2011.03.005 PMID: 21624784

23. Bertoin, J. The structure of the allelic partition of the total population for Galton-Watson processes with neutral mutations. The Annals of Probability. 2009;p. 1502–1523.

24. Chaumont, L, Liu, R. Coding multitype forests: application to the law of the total population of branching forests. Transactions of the American Mathematical Society. 2014;(in press).

25. Kucharski AJ, Kwok KO, Wei VWI, Cowling BJ, Read JM, Lessler J, et al. The Contribution of Social Behaviour to the Transmission of Influenza A in a Human Population. PLoS Pathog.2014; 10(6): e1004206. doi: 10.1371/journal.ppat.1004206 PMID: 24968312

26. Gog JR, Ballesteros S, Viboud C, Simonsen L, Bjornstad ON, Shaman J, et al. Spatial Transmission of 2009 Pandemic Influenza in the US. PLoS Comput Biol. 2014; 10(6):e1003635. doi: 10.1371/journal.pcbi.1003635 PMID: 24921923

27. Kucharski AJ, Edmunds WJ. Cross-immunity and age patterns of influenza A(H5N1) infection. Epidemiology and Infection. 2014; 13:1–6.

28. Bai T, Zhou J, Shu Y. Serologic Study for Influenza A (H7N9) among High-Risk Groups in China. New England Journal of Medicine. 2013; 368(24):2339–40. doi: 10.1056/NEJMc1305865 PMID: 23718151

29. Wang W, Peng H, Tao Q, Zhao X, Tang H, Tang Z, et al. Serologic assay for avian-origin influenza A (H7N9) virus in adults of Shanghai, Guangzhou and Yunnan, China. Journal of Clinical Virology.2014; 60(3):305–8. doi: 10.1016/j.jcv.2014.04.006 PMID: 24793969

30. Meyer B, Drosten C, Müller MA. Serological assays for emerging coronaviruses: Challenges and pitfalls. Virus Research. 2014; 194:175–83.

31. Arinaminpathy N, McLean A. Evolution and emergence of novel human infections. Proceedings of the Royal Society B. 2009; 276 (1675):3937–43. doi: 10.1098/rspb.2009.1059 PMID: 19692402

32. Melegaro A, Jit M, Gay N, Zagheni E, Edmunds WJ. What types of contacts are important for the spread of infections?: Using contact survey data to explore European mixing patterns. Epidemics.2011; 3(3–4):143–51. doi: 10.1016/j.epidem.2011.04.001 PMID: 22094337

33. World Health Organisation. Disease Outbreak News. whoint/csr/disease/ebola/en/. 2014;.

34. Fine P, Jezek Z, Grab B, Dixon H. The transmission potential of monkeypox virus in human populations. International journal of epidemiology. 1988; 17(3):643–650. doi: 10.1093/ije/17.3.643 PMID: 2850277

35. Débarre F, Bonhoeffer S, Regoes RR. The effect of population structure on the emergence of drug resistance during influenza pandemics. JR Soc Interface. 2007; 4:893–906. doi: 10.1098/rsif.2007.1126

36. Antia R, Regoes RR, Koella JC, Bergstrom CT. The role of evolution in the emergence of infectious diseases. Nature. 2003; 426:658–661. doi: 10.1038/nature02104 PMID: 14668863

37. Cooper BS, Pitman RJ, Edmunds WJ, Gay NJ. Delaying the international spread of pandemic influenza. PLoS Med. 2006; 3(6):e212. doi: 10.1371/journal.pmed.0030212 PMID: 16640458

38. Read JM, Lessler J, Riley S, Wang S, Tan LJ, Kwok KO, et al. Social mixing patterns in rural and urban areas of southern China. Proceedings of the Royal Society B: Biological Sciences. 2014; 281 (1785):20140268. doi: 10.1098/rspb.2014.0268 PMID: 24789897

39. Horby P, Pham QT, Hens N, Nguyen TTY, Le QM, Dang DT, et al. Social contact patterns in Vietnam and implications for the control of infectious diseases. PLoS One. 2011; 6(2):e16965. doi: 10.1371/journal.pone.0016965 PMID: 21347264

40. Diekmann O, Heesterbeek J, Metz J. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. Journal of mathematical biology. 1990; 28(4):365–382. doi: 10.1007/BF00178324 PMID: 2117040

41. Gilks, WR, Richardson, S, Spiegelhalter, DJ. Markov chain Monte Carlo in practice. Chapman & Hall/CRC; 1996.

42. Burnham K, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. 2nd ed. Springer-Verlag, New York; 2002.

43. Fiebig L, Soyka J, Buda S, Buchholz U, Dehnert M, Haas W. Avian influenza A (H5N1) in humans: new insights from a line list of World Health Organization confirmed cases, September 2006 to August 2010 Euro Surveill. 2011; 16:32.

44. Kucharski AJ, Mills HL, Pinsent A, Fraser C, Van Kerkhove MD, Donnelly CA, et al. Distinguishing between reservoir exposure and human-to-human transmission for emerging pathogens using case onset data. PLoS Currents Outbreaks. 2014; 7(6).

45. Chowell G, Simonsen L, Towers S, Miller MA, Viboud C. Transmission potential of influenza A/H7N9, February to May 2013, China. BMC Med. 2013; 11(214).