Infectious Disease Modelling 5 (2020) 409–441

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

Using statistics and mathematical modelling to understand infectious disease outbreaks: COVID-19 as an example

Christopher E. Overton [a, k, 1, *], Helena B. Stage [a, 1, **], Shazaad Ahmad [f, m], Jacob Curran-Sebastian [a], Paul Dark [e, n], Rajenki Das [a], Elizabeth Fearon [c], Timothy Felton [e, i], Martyn Fyles [a, l], Nick Gent [d], Ian Hall [a, d], Thomas House [a, b], Hugo Lewkowicz [j, a], Xiaoxi Pang [a], Lorenzo Pellis [a], Robert Sawko [b], Andrew Ustianowski [g, h], Bindu Vekaria [a], Luke Webb [a]

[a] Department of Mathematics, University of Manchester, UK
[b] IBM Research, Hartree Centre, SciTech Daresbury, UK
[c] Department of Global Health and Development, London School of Hygiene and Tropical Medicine, UK
[d] Emergency Response Department, Public Health England, UK
[e] Division of Infection, Immunity and Respiratory Medicine, NIHR Biomedical Research Centre, University of Manchester, UK
[f] Department of Virology, Manchester Medical Microbiology Partnership, Manchester Foundation Trust, UK
[g] Regional Infectious Diseases Unit, North Manchester General Hospital, UK
[h] School of Medical Sciences, University of Manchester, UK
[i] Intensive Care Unit, Wythenshawe Hospital, Manchester University NHS Foundation Trust, UK
[j] Department of Health Sciences, University of Manchester, UK
[k] Department of Mathematical Sciences, University of Liverpool, UK
[l] The Alan Turing Institute, UK
[m] Manchester Academic Health Sciences Centre, UK
[n] Critical Care Unit, Salford Royal Hospital, Northern Care Alliance NHS Group, UK

ARTICLE INFO

Article history:
Received 21 May 2020
Received in revised form 30 June 2020
Accepted 30 June 2020
Available online 4 July 2020
Handling editor: Jianhong Wu

Keywords:
COVID-19
Epidemic modelling
Parameter estimation
Outbreak
Bias
Intervention

ABSTRACT

During an infectious disease outbreak, biases in the data and complexities of the underlying dynamics pose significant challenges in mathematically modelling the outbreak and designing policy. Motivated by the ongoing response to COVID-19, we provide a toolkit of statistical and mathematical models beyond the simple SIR-type differential equation models for analysing the early stages of an outbreak and assessing interventions. In particular, we focus on parameter estimation in the presence of known biases in the data, and the effect of non-pharmaceutical interventions in enclosed subpopulations, such as households and care homes. We illustrate these methods by applying them to the COVID-19 pandemic.

* Corresponding author. Department of Mathematics, University of Manchester, UK.
** Corresponding author.
E-mail addresses: christopher.overton@manchester.ac.uk (C.E. Overton), helena.stage@manchester.ac.uk (H.B. Stage).
Peer review under responsibility of KeAi Communications Co., Ltd.
[1] These authors contributed equally to this paper.

https://doi.org/10.1016/j.idm.2020.06.008

1. Introduction

Mathematical epidemiology is a well-developed field. Since the pioneering work of Ross in malaria modelling (Ross, 1910) and Kermack and McKendrick's general epidemic models (Kermack & McKendrick, 1927), there has been gathering interest in using mathematical tools to investigate infectious diseases. The allure is clear, since mathematical models can provide powerful insight into how these complex systems behave, which in turn can enable these problems to be better controlled/prevented.

Not only is the power of the mathematical tools increasing, but the availability of data on infectious diseases, whether this be a rapid release of data during an outbreak or detailed collection of data for endemic pathogens, is increasing. Rapid interpretation of epidemiological data is critical for the development of effective containment, suppression and mitigation interventions, but there are many difficulties to interpreting case data in real-time. These include interpreting symptom progression and fatality ratios with delay distributions and right-censoring, exacerbated by exponential growth in cases leading to the majority of case data being on recently infected individuals; lack of clarity and consistency in denominators; inconsistency of case definitions over time and the eventual impact of interventions and changes to behaviour on transmission dynamics. Mathematical and statistical techniques can help overcome some of these challenges to interpretation, aiding in the development of intervention strategies and management of care. Examining key epidemiological quantities alongside each other in a transmission model can provide quantitative insights into the outbreak, testing the potential impact of intervention strategies and predicting the risk posed to the human (or animal) host population and healthcare preparedness.

Mathematical modelling has been used as part of the planning process during outbreak response by governments worldwide for many recent outbreaks. For example the UK Department of Health has a long established committee Scientific Pandemic Influenza group on Modelling, or SPI-M to advise on new and emerging respiratory infections (Department of Health and Social Care, 2018). One of the largest instances of such an outbreak in recent history was the 2009 H1N1 pandemic. The World Health Organisation developed a network of modelling groups and public health experts to work on exploring various characteristic of the outbreak (Biggerstaff et al., 2020; Van Kerkhove & Ferguson, 2012). These ranged from characterising the dynamics of the outbreak to investigating the effectiveness of different intervention strategies. This integration of mathematics into policy design indicates the important insights that modelling and statistics can provide.

This paper is a collection of work-streams addressing various technical questions faced by the group as part of the ongoing response to COVID-19, and as such is written to be reflective of the experience we have gone and are currently going through. Therefore, to aid the reader each section includes results and a short discussion. Many of the questions and techniques presented here can be further developed as the availability of data and research interests evolves, but are compiled into this manuscript as an overview of methodology and scientific approaches beyond the standard SIR textbook model that benefit the ongoing efforts in tackling this and other outbreaks.

1.1. COVID-19 pandemic background

First documented in December 2019, an outbreak of community-acquired pneumonia began in Wuhan, Hubei Province, China. In January, this outbreak was attributed to a novel coronavirus, SARS-CoV-2. The initial spread of the pathogen in Wuhan was fast, and after a period of case-finding and contact tracing, China moved to implement a 'shutdown' of Wuhan on January 23, and other cities in China the following days, to try to suppress the growth of the epidemic. These measures may have succeeded at slowing down the rate at which cases have been seeded elsewhere, but in many countries initial importation of cases and transmission has not been contained. Countries around the world are now seeing outbreaks that are overwhelming, or have the potential to overwhelm, healthcare systems and cause a high number of deaths even in high-income countries (Remuzzi & Remuzzi, 2020).

While the majority of documented symptomatic cases are mild, characterised in many reports by a persistent cough and fever, a significant proportion of these individuals go on to develop pneumonia, with some then developing acute respiratory failure and a small proportion of overall cases becoming fatal. Severity of symptoms has been observed to increase with age and with the presence of underlying health conditions such as diabetes (Fang, Karakiulakis, & Roth, 2020) and cardiac conditions, with some evidence that severity of symptoms might depend on gender and ethnicity (Guan et al., 2020; Rimmer, 2020; Wu et al., 2020; Wu & McGoogan, 2020; Zhou et al., 2020).

SARS-CoV-2 has a fast doubling time (the time it takes for the number of cases in the region to double, estimated at approximately 3 days (Pellis et al., 2020) and, potentially, a very large R_0 (the average number of infections caused by each infected individual, with estimates ranging from 1.4 to 6.47 (Liu, Gayle, Wilder-Smith, & Rocklöv, 2020; Mahase, 2020; Majumder & Mandl, 2020 ; World Health Organisation, 2020)). It is possible that there is a significant degree of asymptomatic and/or pre-symptomatic transmission (Li et al., 2020; Mizumoto, Kagaya, Zarebski, & Chowell, 2020; Nishiura et al., 2020), though without robust serosurveys, this is difficult to quantify with certainty. These characteristics result in the pathogen being able to spread widely, rapidly and undetected, presenting a significant risk to public health.

Typically, the aim of an intervention strategy would be to push and keep the reproduction number R_t, defined as the average number of cases generated by a typical infective at time t, below 1. At this point each infected individual subsequently infects, on average, less than one individual, such that the number of cases should decline. The basic reproduction number, R_0,

represents the initial value of R_t, before any intervention is put in place and the population can be assumed to be fully susceptible.

High R_0, fast growth, and possible pre- or asymptomatic infection make the design of potential interventions, and the modelling that would inform them, particularly challenging. Large values of R_0 mean a substantial amount of transmission needs to be halted; fast growth causes the number of cases in the absence of interventions to rise rapidly, so that the time scale of interventions to reduce R_0 must also be fast in order to effect substantive early changes on a population level; finally, the resulting interventions must encompass possible pre- and asymptomatic cases, a challenging prospect when in many instances these individuals are indistinguishable from healthy individuals. Consequently, we must consider the possibility of interventions that are massively disruptive to society and may have to be sustained for a long period of time in order to cause the number of infections to decline towards zero (Ferguson et al., 2020). If infections remain, and the susceptible proportion of the population remains above the herd immunity threshold, these interventions must be upheld to prevent a second wave of the epidemic. There is not yet conclusive evidence as to the degree and duration of immunity conferred by infection with SARS-CoV-2 nor the feasibility of a vaccine, the timeline for which is unlikely to be any time shorter than 18 months away at the time of writing (Jack, 2020). Therefore, short term extreme interventions are not as effective as they might be in other circumstances, since after their removal there remains a long period of time in which cases can rise again. The longer these significantly suppressive and disruptive interventions are in effect, the more severe the effect on the economy, and broader societal health and well-being. Furthermore, adherence to interventions will likely vary with their duration and severity.

We are further challenged by the lack of transferable intuition. Early work looked at intuition gained from SARS and MERS outbreaks, also caused by coronaviruses. Some parameters do appear to be similar to these pathogens, such as the average length of the incubation period (Lauer et al., 2020; Varia et al., 2003; Virlogeux, Fang, Park, Wu, & Cowling, 2016). However, there are also clear differences, with both SARS and MERS being more fatal, but seemingly less efficient at spreading since they did not seed major global pandemics. Another complication is the spread of the infection during the Chinese Spring Festival, a time period during which movement, social, and contact patterns vary significantly. This presents significant challenges as experience and intuition from other studies regarding population mixing and spatial patterns must either be modified or are invalid. Furthermore, the pandemic has received a proportionately larger level of public attention than e.g. the 2009 H1N1 pandemic (Chew & Eysenbach, 2009; Rubin, Amlôt, Page, & Wessely, 2009), largely boosted by social media. This greater level of public awareness, and the successive, staggered interventions placed to prevent disease spread are responsible for significant variations in behaviour (Butler, 2014; Funk, Gilad, Watkins, & Jansen, 2009) and adherence to public guidance both in China and abroad.

The structure of this paper follows two main themes. In Section 2, we discuss various biases that are present in outbreak data and techniques for estimating epidemiological parameters. Accounting for biases and producing robust parameter estimates is important throughout the duration of an epidemic, both for increasing our understanding of the underlying dynamics, and for feeding into models. Firstly, we discuss a bias-corrected method for estimating the incubation period, which can also be applied to serial intervals, onset-to-death time, and other delay distributions. We then present a method for estimating the true growth rate of the epidemic, accounting for the bias encountered since infected individuals may be exported from the region. Our next method is a tool for estimating the expected size of the next generation of infectives based on the rate of observed cases. This tool provides insight into the size of small outbreaks, which can inform decision making when trying to prevent a major outbreak taking off.

In Section 3, we propose a variety of mathematical models looking at disease impact and intervention strategies, with particular focus on non-pharmaceutical interventions due to the current lack of widely deployable, targeted pharmaceutical treatments. These models focus on enclosed populations, since this is the level at which most interventions are implemented. Since the disease is particularly fatal in the elderly and other at-risk groups, we develop a care home model to investigate how the pathogen may spread through care homes. We also develop household models to investigate the impact of different intervention/control strategies. These models can inform policy design for mitigating or controlling epidemic spread. Finally, in the context of relaxing strong social distancing policies, we investigate the extinction probability of the pathogen. We first consider the extinction probability after lifting restrictions. We then develop a household-based contact tracing model, with which we investigate the extinction probability under weaker isolation policies paired with contact tracing, thus shedding light on possible combinations of interventions that allow us to feasibly manage the infection while minimising the social impact of control policies.

2. Biases and estimation during outbreaks

2.1. Potential biases in the outbreak data

Techniques are constantly developing that enable higher volumes of more accurate data to be collected real-time during an epidemic. These data present a large opportunity for analysis to gain insight into the pathogen and the dynamics of the outbreak. However, although the quality of the data is constantly increasing, there are still many biases present. Some of these are due to the data collection methods, and in an ideal world we would be able to eliminate them, and some are simply due to the nature of the outbreak, and will be present regardless of data collection methods.

During an outbreak, many parameters depend on delay distributions (the length of time between two events), such as the time from infection to symptom onset (the incubation period). If an individual can be followed indefinitely, it is easy to

determine the length of these events. However, in reality only events that occur before a given date are observed. Therefore, the data is subject to censoring and truncation issues. In the incubation period, for example, censoring comes into play since, if we have observed an infection but the individual has not yet developed symptoms, we only have a lower bound on how long it will take them to develop symptoms. To account for this, we can instead condition on observing symptom onset before the cut-off date. However, this leads to a truncation issue, since individuals who were infected close to the cut-off date will only be observed if they have a short incubation period, which leads to an overexpression of short delays.

The number of cases tends to grow exponentially during the early stages of an outbreak, causing the force of infection and the number of reported cases to increase with time. This further complicates the truncation issue since not only are recent cases truncated but they also account for the majority of cases. The growing force of infection also needs to be accounted for, since if the potential time of infection is interval-censored rather than observed directly, the probability that the case was infected in each day of that interval is not constant.

In theory, both of these biases are relatively straightforward to account for. In practice however, there are other biases in the data. One of the major biases is the reporting rate. Although the total number of cases may be reasonably described as growing exponentially with a constant rate in the early stages of an outbreak, high-resolution data may exhibit more complex behaviour. This can be due to a variety of reasons, such as the workload becoming overwhelming, the availability of individual-level data decreasing, the laboratories or offices slowing down activity over the weekend, the case definition changing, the testing capabilities increasing, and so on.

Another uncertainty arises since generally only the date of each event is recorded rather than the time. This presents a large window of uncertainty in the length of the delay, since the time of each event can vary up to 24 h, and for a delay distribution, which depend on two events, it could vary by up to 48 h.

Travel rate is another bias present in the data. For example, this changes the density of observed cases in a region, which can change the apparent growth rate. Intervention strategies present a further bias because this can change the growth rate of the epidemic and the reporting rate. Additionally, estimates of certain parameters may vary depending on the interventions that are implemented, so these need to be considered carefully.

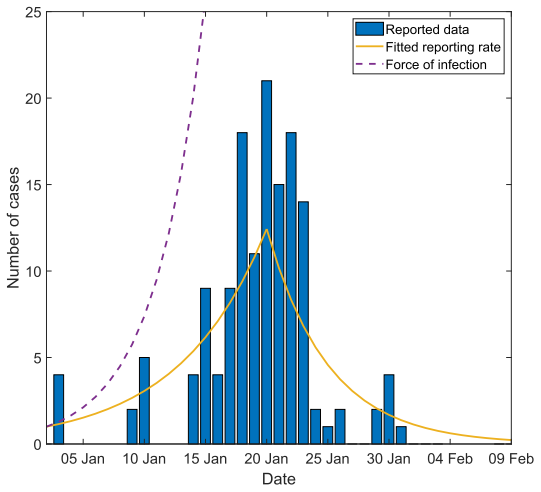
2.2. Incubation period

To model the incubation period, we require information regarding when an individual was infected and when they expressed symptoms. Observing exact time of infection is unlikely, but it can be possible to find potential exposure windows. We consider three different data sets. The first two consist of individuals who travelled from Wuhan before expressing symptoms. We can assume these individuals were infected in Wuhan, since at the time of this data, the force of infection was significantly higher in Wuhan than elsewhere. The length of time spent in Wuhan therefore provides a window during which each individual became infected, and for many of these individuals we also have the date of symptom onset. In the early stages, the growth rate in reported cases was constant, and dependent on the epidemic growth rate in Wuhan and the rate at which people left Wuhan. By using travel to estimate the true number of cases, we estimate the exponential growth rate in Wuhan as $r = 0.25$ (see Section 2.3). Therefore, the force of infection on day i, $g(i)$, is proportional to $e^{0.25i}$. After 23 January when significant travel bans were introduced, the rate at which individuals left Wuhan diminished significantly, causing the reporting rate for our sample dataset to suddenly drop. This occurs since cases are only included if we have a fixed window of time spent in Wuhan prior to developing symptoms. Therefore, if the data is truncated after 23 January, the reporting rate must be appropriately adjusted. This is illustrated in Fig. 1a. The difference between these two datasets is the truncation date, with the first truncated at 20 January and the second at 9 February. The third dataset contains cases that were infected through a discrete infection event, such as spending time with a known infected case. In this "non-Wuhan" dataset, the reporting rate is constant and the force of infection can be assumed constant over each exposure window. The source we use for these three data sets is a publicly available line-list (Sun, Chen, & Viboud, 2020).

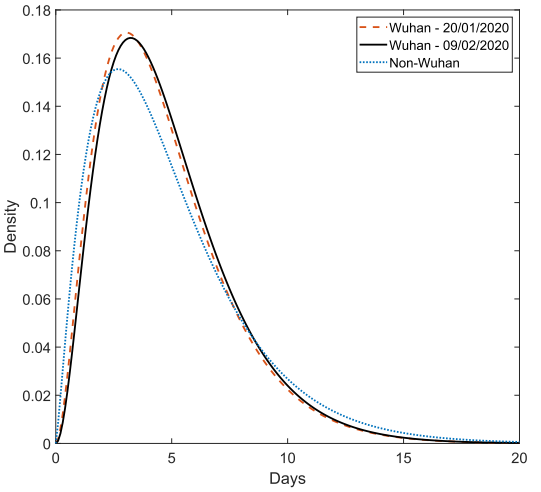
Incubation periods, and many other delay distributions, are generally observed to have right skewed distributions. We therefore choose to use a Gamma distribution, though other distributions can also be applied using the proposed methods, such as Weibull and Log-normal. To fit the data, we use maximum likelihood estimation. To adjust for the biases we use a "forwards" approach (Nishiura, 2010; Scalia Tomba, Svensson, Asikainen, & Giesecke, 2010; Sun, 1995; Svensson, 2007), where we condition on the time of the first event, time of exposure, and find the distribution looking forward to the second event, time of symptom onset. For a data point $\{a_i, b_i, y_i\}$, where infection occurs between a_i and b_i, and y_i is the symptom onset date, the likelihood function is given by

$$L(y_i|a_i, b_i, \theta) = \frac{\int_a^b g(i) f_\theta(y_i - i)\mathrm{d}i}{\int_a^b \int_0^{T-i} g(i) f_\theta(x)\mathrm{d}x\mathrm{d}i},$$

where $g(\cdot)$ is the density function of the infection date and $f_\theta(\cdot)$ is the density function of the incubation period parameterised by θ. From this, the likelihood function for our dataset X is given by



(a) Illustrating the changing reporting rate for the cases being included in our sample as of 9 February. The bars indicate the number of new cases corresponding to the date the individual left Wuhan, and the yellow curve indicates a rough approximation to the reporting rate for the data in the dataset (this is only indicative, since the reporting rate is not necessary in this model). The dotted purple line indicates the force of infection with growth rate $r = 0.25$.



(b) Figure showing the maximum likelihood distributions for the estimated length of the incubation period in days for the three different data sets (with the truncation correction). All three estimates give rise to very similar distributions.

Fig. 1. Reporting rate (a) and maximum likelihood distributions (b) for the COVID-19 incubation period.

$$L(X|\theta) = L(\bigcap_i y_i | \bigcap_i (a_i \cap b_i), \theta) = \prod_i L(y_i|a_i, b_i, \theta).$$

This approach is independent of the reporting rate bias, since the reporting rate depends on the date an individual leaves Wuhan (b_i), which is conditioned against (see Appendix A). We use the mean and standard deviation to characterise the MLE. Since the tail of the incubation period is important when designing quarantine strategies, we then calculate the probability that the incubation period is longer than 14 days and find the minimum day by which 99% of cases will have expressed symptoms (excluding true asymptomatic cases). We also investigate the reporting date uncertainty mentioned in Section 2.1 by considering the different extremes that the data could represent. This is achieved through adding or subtracting a day to all recorded data.

Methods accounting for truncation and growth biases in epidemic data have been discussed widely in the literature (Kalbfleisch & Lawless, 1991; Nishiura, 2010; Su & Wang, 2012; Taylor, Weaver, & Roddy, 2003), however there are fewer applications to outbreaks (Farewell, Herzberg, James, Ho, & Leung, 2005). In the context of COVID-19, estimates have considered growing force of infection, for example (Lauer et al., 2020), and some approaches have considered truncation, for example (Linton et al., 2020). However, these attempts do not adjust for the reporting rate in the data or use the correct force of infection, causing the incubation period to be overestimated. Although the method presented here is independent of the reporting rate, other approaches for estimating the incubation period are not.

2.2.1. *Truncation*

Here we demonstrate the importance of truncation (Table 1). We use the data truncated at 20 January, which has exposure windows between 1 December and 19 January. This data set is chosen since it is most sensitive to truncation due to the exponentially growing force of infection and high reporting rate. Without accounting for truncation, the length of the incubation period is significantly underestimated, which could have a large impact on the success of intervention strategies.

2.2.2. *Different data sets*

To demonstrate the effectiveness of the bias correction method, we compare three different data sets (Table 2). The similar distributions predicted across these datasets suggests a robust method. Fig. 1b compares the full distributions for these three estimates.

2.2.3. *Reporting date uncertainty*

Here we investigate the effect that uncertainty in the reporting date can have on the results, using the data truncated at 9 February (Table 3). The standard interval is the recorded data, wide intervals are obtained by removing a day from the

Table 1
Effect of accounting for truncation on the incubation period.

Method	Mean	Standard deviation	14 day risk	99% confidence date	Sample size
Uncorrected	3.49	2.05	0.00060	10	65
Truncation corrected	4.69	2.78	0.0075	14	65

Table 2
Effect of different data sets on the incubation period.

Method	Mean	Standard deviation	14 day risk	99% confidence date	Sample size
Wuhan - January 20, 2020	4.69	2.78	0.0075	14	65
Wuhan - February 09, 2020	4.84	2.79	0.0081	14	162
Non-Wuhan	4.84	3.22	0.016	16	52

Table 3
Effect of uncertainty in the reporting date on the incubation period.

Method	Mean	Standard deviation	14 day risk	99% confidence date	Sample size
Standard Intervals	4.84	2.79	0.0081	14	162
Wide Intervals	4.21	2.56	0.0041	13	162
Narrow Intervals	5.55	2.86	0.0112	15	162

exposure window lower bound and adding a day to the upper bound, and the narrow interval vice versa. The uncertainty in the reporting date can impact the estimated incubation period, showing that it is important to consider this risk when designing interventions.

2.2.4. Implications

When constructing intervention strategies for an epidemic, the incubation period is an important parameter. For example, consider the quarantine strategy deployed in many countries during the early stages of the epidemic, aimed at preventing cases being imported from Wuhan. This strategy quarantined individuals upon their return from Wuhan for 14 days. For such a strategy to be effective, we require most incubation periods to be less than 14 days, so that the majority of infected people would develop symptoms before quarantine ended, enabling them to be further isolated. In this analysis, we show that in the worst-case scenario we would expect 1 in 62 cases to slip through this quarantine, with the best fit predicting 1 in 101 cases. Therefore, the 14 day quarantine period would capture the majority of cases. Throughout the epidemic, this seems to have been reasonably successful and prevented early seeding of cases in many countries. However, potentially due to complicated travel patterns or asymptomatic transmission, cases have slipped through detection and not been quarantined, which unfortunately has led to the situation observed today.

In addition to the incubation period, there are many other delay distributions that must be estimated while an epidemic is growing, which can be estimated using the same technique. These include the generation time, the time between two infection events in a transmission chain; the serial interval, the time between symptom onset of an infector to their infectee; and the onset-to-death delay, the time from symptom onset to death.

2.3. Transportation modelling and under-reporting

Transportation modelling plays a crucial role in the early stages of an outbreak; an infected individual may travel outside of the region in which they were originally infected and seed further infections across geographical scales which are impossible to contain. Furthermore, as the rate of travelling increases, the number of observed cases within the known "origin" region decreases, and if exportation is not taken into account this results in an underestimation of the number of cases. These underestimates can be improved by looking at the total number of cases across all known affected regions, but doing so introduces further complications. For example, if an individual has less severe symptoms they may not seek medical assistance, thereby not being recorded as a case at their destination. This underestimation of cases can have significant effects if the traveller is able to infect more people. A new transmission chain can thus be started which remains undetected for some time due to a lacking known connection to the "origin" region.

In the "origin" region an individual with mild symptoms may still be tested for an infection due to a higher level of alertness in the local health care system. However, this level of active case-finding may not be present elsewhere, or may not have been allocated a comparable level of resources. Further complications to this model arise from the incubation period of individuals wherein detection is unlikely, and the variations in movement and mixing between people when preventative measures are put in place.

We consider a metapopulation model seeded with an infection in one of the regions, O, and investigate how exportation from this region combined with variability in case-finding can alter estimates for the doubling time and the expected portion

of the population we expect to identify. This accordingly bounds the proportion of the infected population one would be able to target for personal intervention (e.g. quarantine or treatment). Note that the proportion of identified cases need not necessarily correlate with the proportion of the infected population who exhibit symptoms.

Let us assume that movement from O begins at time $t = t_c$, and occurs with a constant rate ρ; this can be thought of as the surge in travel in China during the beginning of the Spring Festival. In the early phase of an epidemic, we can assume the incidence $I(t) = I_0 e^{rt}$ of cases to be growing exponentially with a rate r. The number of cases at time t which were infected a time τ ago is denoted by $i(t,\tau)$, where the probability of detecting a case that infected a host a time τ ago is given by $p(\tau,\tau_{\text{inc}})$. This probability depends on the incubation period τ_{inc} of the infection, and is decomposed into a detection probability $f(\tau,\tau_{\text{inc}})$ after some time τ which may also depend on the individual's incubation period, and the probability density function of said incubation period, $g(\tau_{\text{inc}})$. Hence, $p(\tau,\tau_{\text{inc}}) = f(\tau,\tau_{\text{inc}})g(\tau_{\text{inc}})$ such that the number of observed cases in O is given by

$$C_O(t) = I_0 e^{-\rho(t-t_c)\Theta(t-t_c)} \int_0^t g(\tau_{\text{inc}}) \int_0^t e^{r(t-\tau)} f(\tau,\tau_{\text{inc}}) \mathrm{d}\tau \mathrm{d}\tau_{\text{inc}},$$

where $\Theta(\cdot)$ is the Heaviside step function, and we have assumed that recovery of cases is negligible over the time scale of case observations. If we consider travel to i other regions from O, the total number of observed cases in all destinations is

$$C_D(t) = \sum_{i \neq O} C_i(t) = \omega C_O(t)\left(e^{\rho(t-t_c)\Theta(t-t_c)} - 1\right),$$

where ω is the mean case-finding ability across all destinations. In the presence of real-time transition probabilities p_{ij} of moving between two regions, these estimates can be further elaborated.

We assume that detection occurs immediately following the end of the incubation period, i.e. $f(\tau,\tau_{\text{inc}}) = \Theta(\tau - \tau_{\text{inc}})$. Similarly, we assume a gamma-distributed incubation period $g(\tau_{\text{inc}}) = \frac{1}{\Gamma(k)\theta^k}\tau_{\text{inc}}^{k-1} e^{-\tau_{\text{inc}}/\theta}$ with shape and scale parameters k and θ, respectively. We can parametrise this distribution using the "non-Wuhan" estimate of the incubation period in Table 2, which yields a gamma-distribution with mean 4.84 and standard deviation 3.22. In contrast to other values in the table, this estimate is obtained from discrete infection events, e.g. contact with a known infected case, and therefore has a constant reporting rate, and a constant force of infection over each exposure window. Therefore, this estimate of the incubation period does not rely on the exponential growth rate unlike other estimates from Section 2.2.

Historic estimates for Chinese travel data indicate a mean travel rate from Hubei province of $\rho = 0.029$ which began on January 10 (Keju, 2019; Morris). Using the above incubation period, this suggests a rate $r = 0.22 \pm 0.01$ when ignoring travel exportation, in contrast to $r = 0.25 \pm 0.01$ when accounting for ρ. This difference may seem small, but it reduces the doubling time by approximately 12 h. The expected value of r grows linearly with the exportation rate, which has also been observed with real-time travel models (Kraemer et al., 2020). Further models have also been developed which consider travel and exportation of cases in greater detail (Chinazzi et al., 2020; Gostic, Gomez, Mummah, Kucharski, & Lloyd-Smith, 2020).

The relationship between the observed cases in our origin and destinations can be used to determine the case-finding ability, though it should be noted that ω likely varies with time as burdens are increased on public services and the number of cases grow. Early estimates using data from (Recorded daily case updat, 2020) indicate at most an 80% case-finding ability, suggesting thousands of undetected cases exported to other regions of China, a sufficient quantity to sustain further transmission post-exportation independently of the number of asymptomatic cases present.

The intention of these estimates is not to provide specific values for the doubling time of the spread of COVID-19 in China (as the estimates above use historic travel data and are limited by the availability of data), but to bring attention to the unusual circumstances surrounding changes in contact patterns, and mobility during the Chinese Spring Festival, the largest human migration on Earth (Keju, 2019). Failing to account for the significant level of dispersion or exportation of cases during these circumstances will significantly skew our estimates.

2.4. Estimating the size of the first generation from the observed number of symptomatic individuals

In a scenario where a single individual exposes a group to infection, it can be unclear how many people have been infected since they do not immediately develop symptoms. However, knowing the true prevalence in the population is essential to determine the most effective interventions to put in place, and to estimate future burdens on public services. Using the probability density function of the incubation period, we consider the efficacy of using the time it takes for people to present with symptoms as a predictor for the size of the infected group. This analysis is an effective ready reckoner at early stages of a novel infection, or in close contact environments, and is useful for predicting generation size when a complete data set is not yet available. In this analysis we focus on a scenario where infection time is known. In reality, we may only know an exposure window. For short exposure windows this method can still be valid, but for longer exposure windows it will need extending to account for this added uncertainty.

We assume that the number of individuals who have been exposed to potential infection is known, in which case the number of people who are infected can be assumed to be binomially distributed with an unknown probability P that each

individual has been infected. To determine the distribution of infected individuals, we use the available information regarding the number of individuals who have expressed symptoms. This yields two cases. In the first case, we assume that the true number of symptomatic individuals are observed. In the second case, we take the number of observed symptomatic individuals as a lower bound on the true value.

We wish to determine the probability that the first generation has e_0 individuals, $E_0 = e_0$, given that i_τ symptomatic individuals are observed on day τ, $I_\tau = i_\tau$. This is given by (see Appendix B)

$$\mathbb{P}(E_0 = e_o | I_\tau = i_\tau) = \frac{e_0!}{(e_0 - i_\tau)!}(1 - F(\tau))^{e_0} \times \frac{(1 - F(\tau))^{-i_\tau}(n - i_\tau)!(i_\tau + 1)}{(n+1)!\,{}_2F_1(i_\tau + 1, i_\tau - n, i_\tau + 2, F(\tau))}.$$

This gives a distribution of the generation-size based on the number of observed symptomatic individuals by time τ. We can extend it to investigate a scenario where no symptomatic individuals have been observed by time τ by using a value of 0 for I_τ:

$$\mathbb{P}(E_0 = e_0 | I_\tau = 0) = \frac{(1 - F(\tau))^{e_0}}{(n+1)\,{}_2F_1(1, -n, 2, F(\tau))}$$

This can be used to illustrate worst and best case scenarios given τ time has passed without symptomatic individuals. Additionally, if we consider the probability that $E_0 = 0$, we can find the value of τ where we can have a 95% confidence that there will not be a second generation:

$$\mathbb{P}(E_0 = 0 | I_\tau = 0) = \frac{1}{(n+1)\,{}_2F_1(1, -n, 2, F(\tau))} > 0.95$$

This analysis considers the case when the number of observed symptomatic individuals to date is the true number. In practice however, we do not generally observe every symptomatic individual, so the number of observations is only a lower bound on the true number. To address this, rather than considering I_τ as the total number of people who have developed symptoms by time τ, we can define $\tilde{I}_\tau$ as the minimum number of people who have developed symptoms by time τ. We assume that the probability that $\tilde{I}_\tau$ is equal to $\tilde{i}_\tau$ for a given value of i_τ is uniform at $\frac{1}{i_\tau + 1}$. We can then use the same methods as above to infer a distribution for P. Details are provided in Appendix C.

As we can see from Fig. 2, this method can be used to predict the number of infected individuals in the original exposed group. However, we have also demonstrated the importance of caution when interpreting this data. If there is uncertainty surrounding the presentation of symptomatic patients, using $\tilde{I}_\tau$ as a lower bound is a robust method to ensure the size of the generation is not underestimated.

3. Modelling intervention strategies

3.1. Adherence

When designing intervention strategies, we need to consider how adherence may alter their effectiveness. This is important, since highly effective interventions may not be adhered to if they present great individual cost to a population. In this case, a theoretically less effective intervention may perform better, if it has sufficient reduction in individual-level cost. In this section, we illustrate the potential impacts of adherence on the effectiveness of interventions using a toy model.

Consider a standard SIR model, and denote by $S(t)$ and $R(t)$, respectively, the susceptible and recovered/immune fractions of the population at time t. We can write S in terms of R such that

$$S(t) = S(0)\exp(-R_0 R(t))$$

and let $t \to \infty$ to get the final size formula

$$1 - R(\infty) = S(0)\exp(-R_0 R(\infty)),$$

where $R(\infty)$ is the fraction of cases at end of outbreak in the absence of behavioural change (Brauer, 2019). This gives a ready reckoner for the eventual attack rate if interventions are not put in place or come in too late to be effective. To illustrate, if we have $R_0 = 3$ (and $S(0) \approx 1$), then $R(\infty) = 0.94$. If an intervention is put in place that reduces (with full adherence) $R_0 < 1$ then the outbreak will be controlled. Indeed, let us assume that R_0 is reduced to zero by the intervention: for example, assume that social distancing is perfect and the number of contacts of a fully-adherent individual is zero. If only 50% of people adhere to the intervention then the average number of contacts is effectively reduced by a half and logically $R_0^\dagger = R_0/2 = 1.5$ (the $\dagger$ representing quantities post intervention) and $R(\infty)^\dagger = 0.58$ in this case. However, this assumes that adherence is an independent random process at each contact. This suggests that for each contact an individual would ordinarily make, they "toss a coin" to decide whether to isolate or not. In reality, individuals are more likely to show polarity, where some individuals

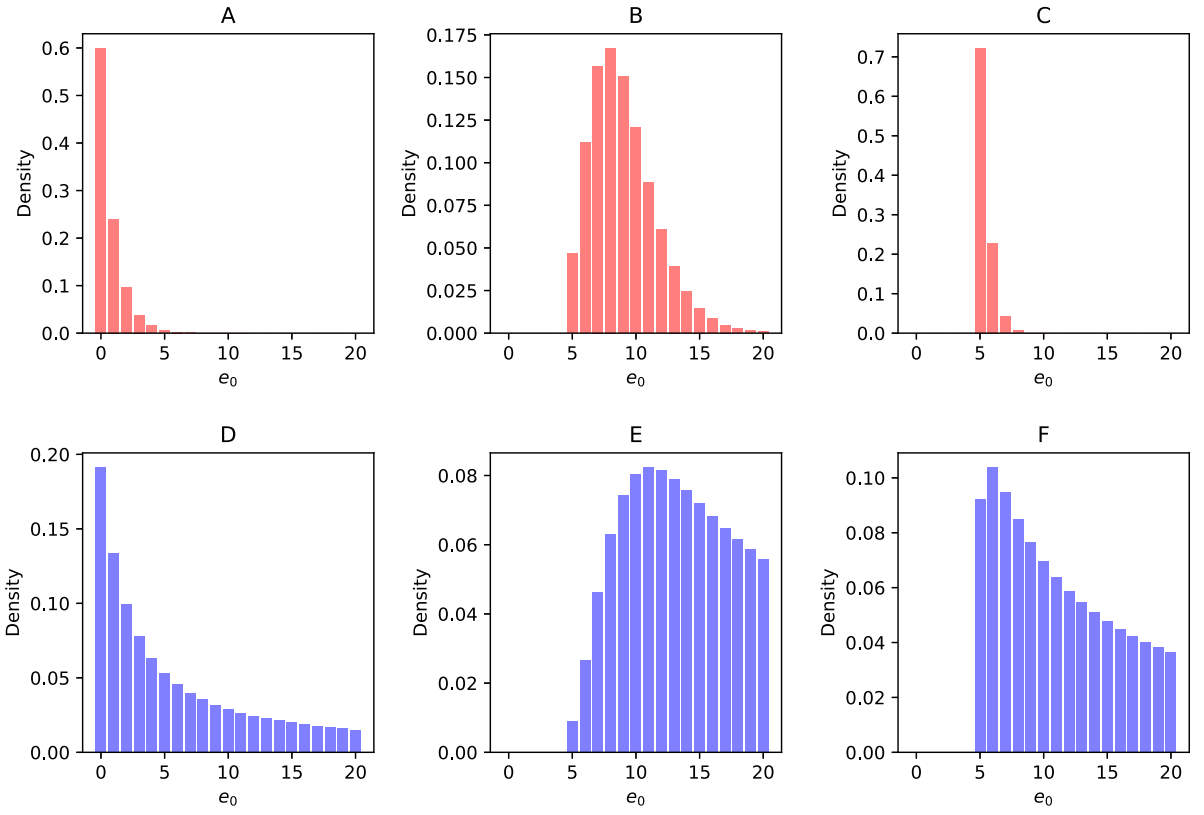


Fig. 2. Prediction of the size of the first generation, e_0, in an infection event in which 20 people were exposed. A, B and C show the density when the number of observed symptomatics is taken to be the true number of symptomatics, D, E and F consider the case where the observed symptomatics is a lower bound on the true symptomatics. A and D consider the case when zero symptomatics are observed after 5 days, B and E when 5 are observed after 5 days, and C and F when 5 are observed after 10 days. The incubation period for the disease has been modelled as a gamma distribution with a mean of 4.84 and standard deviation of 2.79 (Table 2).

reduce all their contacts and follow the measures and a proportion of individuals choose to not adhere to the intervention. If there was distinct polarity in the population such that 50% adhered perfectly and 50% ignored policy, then a toy model can be created with two infectious groups, I_A and I_B, that behave differently. In this case

$$\dot{S} = -(R_A I_A + R_B I_B)S,$$

$$\dot{I}_A = \varphi(R_A I_A + R_B I_B)S - I_A,$$

$$\dot{I}_B = (1-\varphi)(R_A I_A + R_B I_B)S - I_B,$$

where a dot over a variable represents its time derivative. Such an epidemic model, where the two groups have the same susceptibility but different infectivity, has the same final size as an epidemic in a single-type model with the same R_0 (e.g. see (Andreasen, 2011)). However, they have different durations as can be seen in Fig. 3, where $\varphi = 1/2$, $R_B = 0$ and $R_A = 3$. This shows that the assumptions about the nature of adherence predict the same growth rate and final size, but that the more polarised adherence has faster early growth and therefore an earlier peak.

More complicated model structures could be constructed by incorporating adherence with intervention by susceptible states, which would lead to core group dynamics (see for example (Keeling & Rohani, 2011)). This issue of independent versus polarised adherence is related to the idea of all-or-nothing versus leaky vaccination (Goldstein et al., 2009; Magpantay, Riolo, Domenech de Cellès, King, & Rohani, 2014), where you either vaccinate a fraction of the population with 100% efficacy or vaccinate 100% of the population with reduced efficacy (House & Keeling, 2011). Note however that vaccination reduces your susceptibility (whether only or also), rather than only your infectivity as in the model discussed above, and variation in

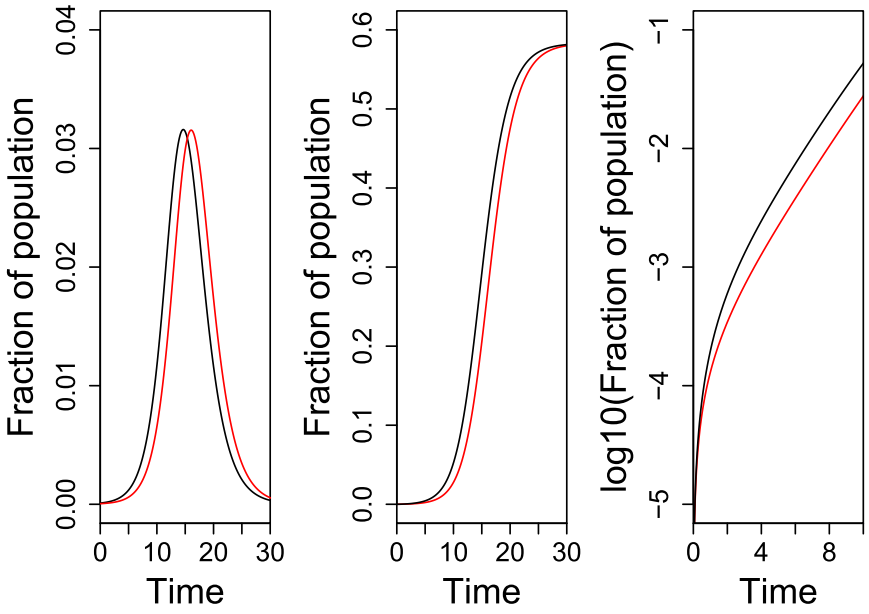


Fig. 3. Comparing the different definitions of adherence. The left panel shows the people in I state from SIR model (red) and from I_A state plus the I_B state in SI_AI_BR model (black), with the same overall mean level of adherence. The resulting recovered curves are in the middle panel, with the right panel showing the recovered cases on log scale.

susceptibility does reduce the final size, with imperfect coverage with a perfect vaccine (all-or-nothing) leading to a lower final size than full coverage with a leaky vaccine (all individuals having the same mean susceptibility).

3.2. Care home model

The ongoing COVID-19 outbreak is known to have higher mortality rates amongst the elderly, the immunocompromised and those with respiratory and health complications (Guan et al., 2020; Wu et al., 2020; Wu & McGoogan, 2020; Yang et al., 2020; Zhou et al., 2020). In this section, we model the introduction of an infectious disease into care homes, in order to obtain estimates of the final size of the epidemic in the vulnerable population as well as predictions for the number of hospitalisations and fatalities.

Modelling of care homes in the UK is conducted against the backdrop of a wider epidemic in the general population, which we here assume to be following SEIR dynamics with a basic reproduction number R_0 that might be different from the within-care home reproduction number R_C.

Care homes are assumed to be closed populations, with the infection entering each of them independently with a certain probability. Infection is seeded only once, and within-care home outbreaks then evolve independently from, and do not contribute to, other care home outbreaks and the epidemic in the background population. To keep track of hospitalisations, we model the within-care home infection dynamics using a compartmental model that, in addition to SEIR model, has compartments for mildly symptomatic prodromal cases (P), who show no symptoms but are capable of transmitting the virus, those who recover from the disease after mild symptoms that did not require hospitalisation (M), those who have severe symptoms and are admitted to hospital (H), those who recover after hospitalisation (R), and those that die (D). This is illustrated in Fig. 4.

The stochastic component of the model, i.e. the random introduction of the infection in care homes, is modelled using the Sellke construction (Andersson & Britton, 2000). Each care home i is given an individual, random threshold of resistance, Q_i, which is drawn from an $\mathrm{U}(0,1)$ distribution. At time t, we then calculate the infection pressure $\mathrm{IP}(t)$ from the background epidemic so that care home i becomes infected at time T_i, where $T_i = \inf\{t|\mathrm{IP}(t) > Q_i\}$. The infection pressure up to time t for a median sized care home is the integral from 0 to t of the force-of-infection (FOI) applied to the care home coming from all infectious sources, multiplied by a probability p. This probability represents the probability of the infection being introduced to a median-sized care home. For other care homes, we allow this probability to be proportional to its size, under the assumption that larger care homes employ more staff and are therefore at higher risk of introduction. When the infection pressure becomes higher than an individual care home's resilience threshold, that care home begins its own deterministic infection dynamics with a single initial infected case. The equations describing the background epidemic and the within-care home epidemic are given in Appendix D.

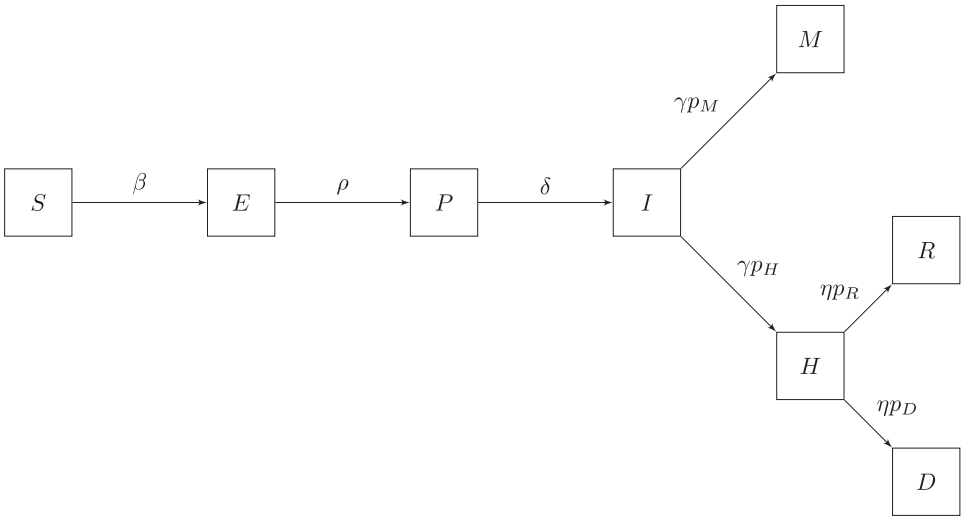


Fig. 4. Compartmental model for disease dynamics within a care-home. We extend a deterministic SEIR model to include compartments for prodromal (infectious) cases (P), mildly symptomatic cases that recover without requiring hospitalisation (M), cases that do require hospitalisation and are removed from the care home (H), cases that die in hospital (D) and cases that recover in hospital (R).

We run this model on data for the entire care home population in the UK, so that there are approximately 15,000 care homes with a total population of approximately 450,000 residents (Care Quality Commission, 2020). Care home sizes range from 1 to 215, with a mean size of 29.4. In this model we only consider the vulnerable population within care homes. We assume $R_0 = 1.5$ in the background epidemic, a relatively low value that somehow accounts for a certain degree of control, and an $R_C = 3$ to allow relatively explosive epidemics in care homes due to potentially more frail individuals, difficulty in isolation and staff inadvertently passing the infection from one case to the next. The other parameters in the baseline scenario are reported in Table 4. Apart from the reproductive number, the background epidemic uses the same parameters as the care home epidemic. However, in the background model there are only rates from E to I and I to R, which are taken to be the rates from E to P and I to M, respectively.

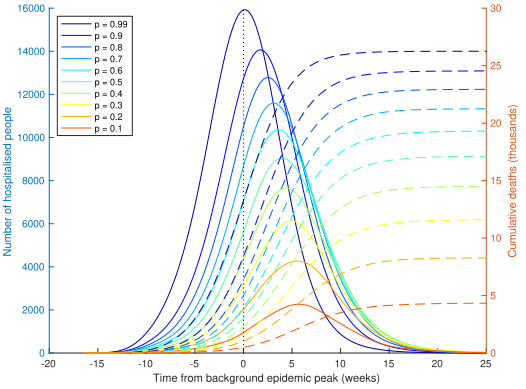
Fig. 5a shows number of hospital beds occupied and the cumulative number of deaths for the parameter values chosen and for different values of p. Time is shown in weeks, where week zero represents the peak of the external/background epidemic. Fig. 5b summarises the first, showing the impact of reducing p on the demand for hospital beds and on the final number of deaths. It also shows the impact that changing p has on the timing of the peak.

Reducing p corresponds to increasing protection of the vulnerable population in care homes by preventing introduction of infection (e.g. screening, testing and promoting hygiene among staff, etc.), a policy sometimes termed 'cocooning' or 'shielding'. The results show that reducing p from 0.99 to 0.1 corresponds to a reduction of around 22,000 (83%) in the number of deaths and around 131,500 in the number of hospitalisations (83%). Strategies aimed at reducing the probability of introduction into a care home, such as reducing the number of visitors or increased monitoring and protection of care home staff, are therefore predicted to have a large impact on the number of cases in vulnerable care home populations.

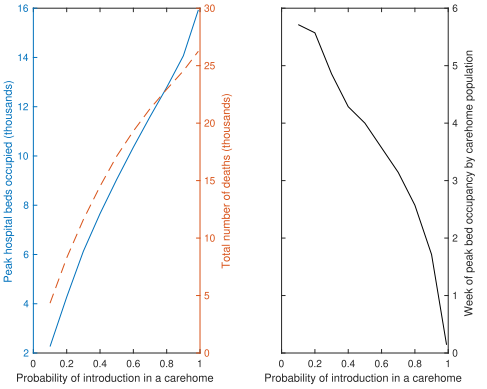
There are a variety of assumptions underpinning this model. Firstly, the background epidemic ignores structure and assumes homogeneous mixing. This is likely to make the peak more pronounced, so presents a worst case scenario for the demand on hospital beds. The assumption of R_0 for the background epidemic only affects the shape and duration of the background epidemic, since the tuneable parameter p controls the risk of introduction to the care home. That is, if R_0 is small, a large p still presents a high force of infection into the care homes. Therefore, for a fixed p, we expect that changing R_0 does not affect the total number of deaths, but it changes the peak hospitalisation incidence because a faster and more explosive background epidemic makes epidemics in care homes more synchronised. In fact, when testing the impact of a longer, flatter background epidemic, for example obtained by simulating three slightly desynchronised background SEIR epidemics, results have lower peaks and a much more variable timing (not shown). Assuming each care home is independent might not be realistic, since it is likely that staff are shared between multiple homes, in which case they can act as vectors of transmission between homes. However, in the model current outbreaks are already quite synchronised (the within care home outbreaks occur at similar times), so the effect of this assumption is likely to be minimal. The final major assumption is that the epidemic within the care homes is deterministic. This removes the probability of random extinction and random delays, and should obviously be relaxed with a stochastic model, given half of care homes have size smaller than 25. However, the extinction probability is very low with $R_C = 3$, so this stochastic effect is unlikely to have a large impact. Random delays, instead, may change the shape and timing of the epidemic, which could potentially reduce the peak burden. Therefore, this model represents a worst case scenario.

Table 4
Information and values for each parameter in our within care home model.

Parameter	Value	Details
R_0	1.5	Basic reproductive ratio (for external epidemic)
R_C	3	Basic reproductive ratio (for care-home epidemic)
B	$R_0 \times \gamma$	Infectiousness
P	1/5	Reciprocal of period between exposure and asymptomatic infectiousness
Δ	1/2	Reciprocal of period between asymptomatic infectiousness and onset of symptoms
Γ	1/4	Reciprocal of infectious period
p_M	0.64	Proportion of vulnerable infectious cases who recover without severe symptoms
p_H	0.36	Proportion of vulnerable infectious cases who are hospitalised
H	1/14	Reciprocal of period of hospitalisation
p_R	5/6	Proportion of hospitalised cases that recover
p_D	1/6	Proportion of hospitalised cases that die
r_P	1/2	Relative infectiousness during prodromal phase



(a) Hospital prevalence (solid lines, left axis) and cumulative number of deaths (dashed lines, right axis). The x-axis shows time in weeks, with 0 (vertical dotted line) denoting the peak of the background epidemic. Colours refer to different values of the probability p that a median-sized care home experiences an introduction.



(b) Left plot shows the predicted height of the peak in the number of hospital beds occupied (solid blue line, left axis) and total number of deaths (dashed red line, right axis), as a function of the probability p of introduction in a median-sized care home. Right plot shows the timing of the peak, which always appears to occur at a similar time or later, compared to the background epidemic. The peak arrives earlier for larger values of p, as high p corresponds to multiple simultaneous introductions early on in the background epidemic.

Fig. 5. Hospitalisation prevalence (a) and hospitalisation peak (b) for the care home model.

3.3. Household isolation modelling

In the absence of cure or vaccine for COVID-19, governments worldwide must rely on non-pharmaceutical interventions (NPIs) to control the outbreak (Haleet al., 2020). A natural such intervention is to ask individuals who express symptoms similar to COVID-19 to isolate themselves, but variants to such individual isolation might include policies sometimes referred to as household isolation, household quarantine and mixed isolation. In this section, we investigate how such strategies affect the spread of the epidemic when bearing in mind that adherence to each intervention may differ.

Individual isolation relies on individuals staying in isolation when they express symptoms, thereby stopping transmission. However, there is potential asymptomatic or prodromal transmission before they go into isolation. Additionally, isolation strategies generally ask infected individuals to remain at home, which presents an infection risk to the other members of their household, who may go on to spread the infection.

The term 'household isolation' refers to a policy where, upon first detection of symptoms within a household, all individuals within the household go into isolation for a fixed duration of time. This strategy reduces the risk that other household members, if they are infected within the household, transmit in the community when pre-symptomatic (and hence before they self-isolate themselves) or if asymptomatic but still infectious.

A blanket policy invoking a fixed duration of household isolation might cover the full epidemic in a small household. However, a larger household might present multiple generations of infection, potentially extending the within-household outbreak beyond the fixed duration of the household isolation policy. To address this issue, 'household quarantine' is another potential strategy. Upon detection of symptoms, the entire household is isolated until a fixed duration of time after

the last symptomatic case within the household expresses symptoms. This ensures that there are no symptomatic cases evading intervention but applies quite drastic measures to the household.

A fourth strategy, that reduces the cost relative to household quarantine, is mixed isolation. Here, upon detection of symptoms the entire household is isolated for a fixed length of time. Any subsequent cases within the household then undergo individual isolation as described above. This reduces the risk of cases not being isolated whilst allowing recovered individuals to return to work. There is however still some remaining risk that infected individuals may not yet express symptoms after the end of the isolation period, but this risk can be controlled through the duration of each isolation.

Although there is now a rich theoretical literature on households models (Ball et al., 2015; Ball, Britton, & Sirl, 2011; Ball, Mollison, & Scalia-Tomba, 1997), the mainstream methodological tools in this research area present important limitations that make them not directly applicable to studying these control policies. First, exact theoretical or asymptotic results in these models are mostly restricted to time-integrated quantities, i.e. those quantities that do not depend on the detailed temporal shape at which the infectivity is spread by an individual: these are R_0 (or any other reproduction number (Ball, Pellis, & Trapman, 2016; Pellis, Ball, & Trapman, 2012), e.g. the household reproduction number R_*), the probability of a large epidemic, and the epidemic final size (Andersson & Britton, 2000). For this reason, the vast majority of the literature relies on the standard stochastic SIR model (Andersson & Britton, 2000), despite its unrealistic infectivity profile. Even if more recent work has expanded beyond time integrated quantities, for example considering the real-time growth rate (Ball et al., 2016; Pellis, Ferguson, & Fraser, 2011), if the interest is on tracking the dynamics of infection spread, a model based on full temporal representation of between- and within-household dynamics (House & Keeling, 2008) appears necessary.

A second limitation of standard household models is the key assumption of constant parameter values. This appears essential for any form of analytical progress. However, in the context of the interventions discussed above, a reduction in transmission between households, as well as a potential increase in within the household, require parameters to change over time.

To overcome these limitations, we consider two approaches. The first approach fully captures both within and between-household dynamics with a master-equation formalism, i.e. by relying on a Markovian within-household dynamics and keeping track of the expected number of households in each possible state of their internal dynamics. The second approach has a greater emphasis on within-household dynamics, and is fundamentally an independent-households, individual-based, stochastic simulation. The more limited mathematical tractability is the price to pay for an increased flexibility, as the within-household Markov assumption is relaxed and exact distributions for delays between events, typically informed by the data, can be explicitly inputted. Although both approaches can account for increased within-household transmission as isolation and quarantine are imposed, we only consider this for the second method here. This aspect allows us to study the increased risk of infection a vulnerable individual in the household would experience following the implementation of a control policy.

To model the households in the UK, we construct a realistic distribution of household sizes (which is given in the supplied code). We take this demographic data from the 2001 Census (Office for National Statistics, 2011). More recent information, though less specific on large household sizes, shows that sizes of smaller households are largely unchanged over time (Office for National Statistics, 2019).

3.3.1. Population and household transmission

In this section, we investigate the above intervention strategies under the assumption that a fraction of households adhere 100% with an intervention and the remaining households ignore the intervention. To model the interventions, we implement a dynamical household model that explicitly represents the small sizes of households.

The dynamics of the outbreak are simulated using an SEPIR model. This model assumes that there are five possible states in which an individual can be. These are, susceptible, latent, mildly symptomatic prodrome, symptomatic infectious and removed. Individuals are infectious during the mildly symptomatic prodrome state and the symptomatic infectious state. Following (Cauchemez, Carrat, Viboud, Valleron, & Boëlle, 2004), we assume that within-household transmission scales with the inverse of the household size to a specified power η. Such a model can be used to investigate how the pathogen spreads through and between households.

The methodology involved is the use of self-consistent differential equations, first written down by Ball (Ball, 1999). More recent developments, including numerical methods for these equations, include (Black, Geard, McCaw, McVernon, & Ross, 2017; House & Keeling, 2008; Kinyanjui et al., 2018; Ross, House, & Keeling, 2010). Important features of this approach include allowing for a small, finite size of each household in which random effects are important and each pair can only participate in one infection event.

Model. Let $Q_{n,s,e,p,i}(t)$ be the proportion of households in the population at time t of size n, with s susceptibles, e exposed, p prodromal, and i symptomatic infectious individuals. The number of recovered individuals will be $n-s-e-p-i$. In the absence of household-based interventions, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}Q_{n,s,e,p,i} = &-\left(sr_{s\to e}(t,\mathbf{Q}) + er_{e\to p} + pr_{p\to i} + ir_{i\to\varphi} + n^{-\eta}sp\tau_p + n^{-\eta}si\tau_i\right)Q_{n,s,e,p,i}\\ &+(s+1)r_{s\to e}(t,\mathbf{Q})Q_{n,s+1,e-1,p,i} + (e+1)r_{e\to p}Q_{n,s,e+1,p-1,i}\\ &+(p+1)r_{p\to i}Q_{n,s,e,p+1,i-1} + (i+1)r_{i\to\varphi}Q_{n,s,e,p,i+1}\end{aligned}$$

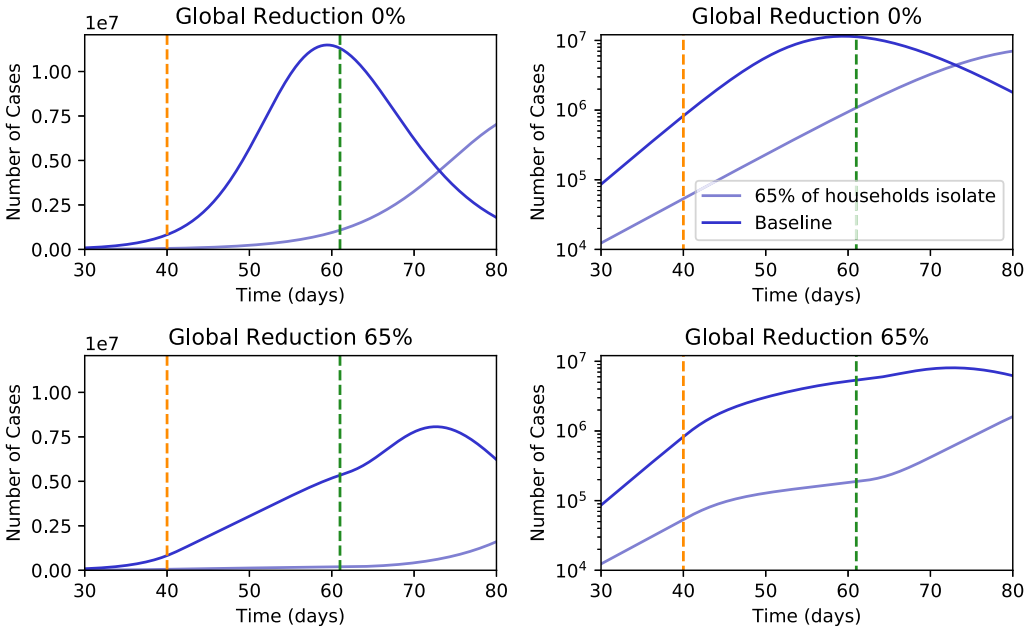


Fig. 6. Investigating the impact of household intervention with $\alpha_W = 65\%$ on the number of cases, for two levels of global intervention. The top two figures have no global intervention, and the bottom two have an $\varepsilon = 65\%$ reduction in global transmission, e.g. school closure or partial lockdown, lasting for 21 days (between the vertical lines). The left-most figures have a linear y-axis. The right-most figures show that same results on a logarithmic y-axis. The household size distribution is taken from the 2001 Census (Office for National Statistics, 2011).

$$+n^{-\eta}(s+1)p\tau_p Q_{n,s+1,e-1,p,i} + n^{-\eta}(s+1)i\tau_i Q_{n,s+1,e-1,p,i} \ ,$$

where we take any Q with logically impossible indices just to equal 0, $r_{a\to b}$ is the rate from state a to b, and τ_a is the transmission rate from an individual in state a. Here $\mathbf{Q}$ is a vector constructed from some ordering (e.g. lexicographic) of the $Q_{n,s,e,p,i}$ (see the code for details). The transmission into households is given by

$$r_{s\to e}(t,\mathbf{Q}) = \Lambda(t) + \sum_{n=1}^{n_{\max}} \sum_{s=0}^{n} \sum_{e=0}^{(n-s)} \sum_{p=0}^{(n-s-e)} \sum_{i=0}^{(n-s-e-p)} \left(p\beta_p(t) + i\beta_i(t)\right) Q_{n,s,e,p,i} \ .$$

Here Λ represents infections imported from outside the population of households, and the other terms represent between-household transmissions. In our code, we assume Λ is a step function. Results are largely insensitive to the precise choice of Λ, but compared to, for example, random seeding of infections in households, starting the whole population susceptible and exposing to a small amount of external infection for a fixed time period has less room for the precise initial condition chosen to influence results, and is more realistic for the situation observed in countries apart from China. We take a 'global' intervention as part of the baseline, in particular, we can model phenomena such a school closures that hold during a set of times $\mathscr{T}$ as

$$\beta_x(t) = \begin{cases} (1-\varepsilon)\beta_x(0) & \text{if } t \in \mathscr{T} \ , \\ \beta_x(0) & \text{otherwise,} \end{cases}$$

for $x \in \{p,i\}$. We call ε the global *reduction*. We will generally drop this t-indexing for simplicity, and will also consider only a household isolation strategy (though the other strategies can be considered similarly, with an example of how other strategies could be captured in this model framework given in Appendix E). Instead of isolating for a fixed duration, we assume that a fraction α_W of households isolates when there is at least one symptomatic case in the household, and isolating households leave isolation when no symptomatic cases remain. We make this assumption since it may potentially capture the behaviour of real households, who are more likely to remain isolated based on presence of symptoms rather than for a fixed duration. In the non-Markovian household model in Section 3.3.2 we consider a fixed duration of isolation as described in the earlier definition. Isolating households do not experience new infections, meaning that the dynamics become

$$\frac{\mathrm{d}}{\mathrm{d}t}Q_{n,s,e,p,i} = -\left((1-\alpha_W 1_{\{i>0\}})sr_{s\to e}(t,\mathbf{Q}) + er_{e\to p} + pr_{p\to i} + ir_{i\to\varnothing} + n^{-\eta}sp\tau_p + n^{-\eta}si\tau_i\right)Q_{n,s,e,p,i}$$
$$+(1-\alpha_W 1_{\{i>0\}})(s+1)r_{s\to e}(t,\mathbf{Q})Q_{n,s+1,e-1,p,i} + (e+1)r_{e\to p}Q_{n,s,e+1,p-1,i}$$

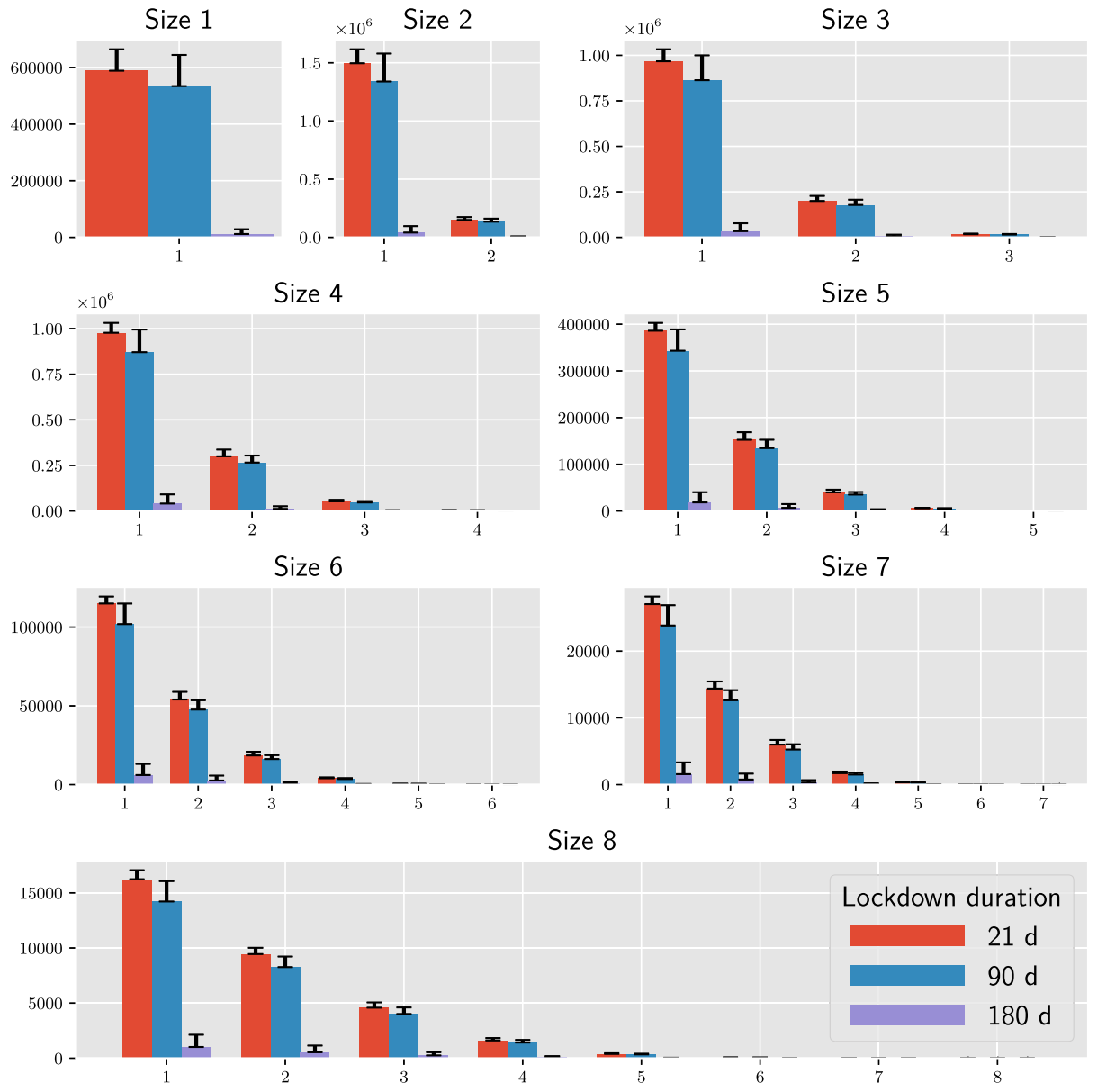


Fig. 7. Histograms representing the number of infectious cases (x-axis) at different household sizes, against a background of household isolation and for three levels of global transmission reduction. The global reduction takes the form of a lockdown reducing contacts by 65% for 21, 90 and 180 days. The number of cases is a cumulative measure of probing the state of each household every two weeks over the 180 day period. The error bar represents a sample standard deviation computed from the simulation outputs ensemble. The ensemble was constructed by sampling uniformly three model inputs: isolation adherence (α_W), global reduction (ε) and secondary probability of attack. Parameters are in the main text.

$$+(p+1)r_{p\to i}Q_{n,s,e,p+1,i-1}+(i+1)r_{i\to\varnothing}Q_{n,s,e,p,i+1}$$
$$+n^{-\eta}(s+1)p\tau_pQ_{n,s+1,e-1,p,i}+n^{-\eta}(s+1)i\tau_iQ_{n,s+1,e-1,p,i}\ ,$$

and also do not transmit outside, meaning that the rate of between-household transmission becomes

$$r_{s\to e}(t,\mathbf{Q})=\Lambda(t)+\sum_{n=1}^{n_{\max}}\sum_{s=0}^{n}\sum_{e=0}^{(n-s)}\sum_{p=0}^{(n-s-e)}\sum_{i=0}^{(n-s-e-p)}(1-\alpha_W 1_{\{i>0\}})(p\beta_p+i\beta_i)Q_{n,s,e,p,i}\ .$$

Parameterisation. Using the methods in (Black et al., 2017; Ross et al., 2010), it is possible to fit household models of this kind to the overall growth rate, r, which we take to correspond to a doubling time of three days. Natural history parameters can then be set directly based on reasonable estimates: $r_{e\to p}$ to the inverse of the latent period; $r_{p\to i}$ to the inverse of the prodromal period; $r_{i\to\varnothing}$ to the inverse of the symptomatic period. Shaw (Shaw, 2016) analyses various household datasets for respiratory pathogens and estimates values for η close to 1, so this is taken to be 0.8. The remaining degrees of freedom are relative infectiousness of the prodrome (taken as a third) and the probability of transmitting within a pair, which we can take as a typical value given by Shaw (Shaw, 2016). For the numerical results in Figs. 6 and 7, the baseline natural history parameters are chosen to be $r_{e\to p} = 1/5$, $r_{p\to i} = 1/3$, $r_{i\to\varnothing} = 1/4$.

Summary. Using the given parameter values for our baseline scenario (Table 5), we consider a combination of household isolation (which follows all-or-nothing adherence) with global reduction in transmission (which follows leaky adherence) for three weeks and show the results in Fig. 6. The distribution of infectious individuals varies with household size, which is shown in Fig. 7 for different durations of global intervention. Applying household isolation at 65% adherence ($\alpha_W = 0.65$) manages to reduce the spread of infection, but appears insufficient in this model and with baseline parameters for controlling the outbreak in the long-term, unless other intervention strategies that reduce the global transmission (increasing ε) are adopted at the same time. Alternatively, different levels of adherence can be considered to determine if and when control may be achieved purely through household-based interventions. For the model proposed in the next section, we look into the effectiveness of increasing adherence.

3.3.2. Non-Markovian models with enhanced within-household transmission

The model described above has the advantage of being able to track the dynamics within the household as well as the overall epidemic in the population in a relatively efficient manner. We now discuss a different framework that loses part of the capability in keeping track of the overall epidemic, but offers further flexibility both in the impact of policies on the within-household dynamics and in the distributions between events in the infectious life of an individual. We use this model to investigate the relative effectiveness of the different control policies. We also consider allowing recovered individuals to leave the household, even in the context of household isolation or household quarantine. This has no impact on the transmission dynamics, but reduces the individuals' life disruption and potential economic cost of any policy implemented.

This model assumes that there is no reintroduction within households so each household can only be isolated or quarantined once. The assumption that only one household member is infected from outside is approximately satisfied if we assume homogeneous mixing between households and a large number of households, which are all fully susceptible at the start of the epidemic. However, the reality of heterogeneous mixing makes reintroduction a likely possibility even early on in the epidemic. This model, therefore, lacks an explicit description of the social network structure beyond the household. For simplicity, we assume that within households all individuals are identical in terms of their disease dynamics, although the method might be extended to allow for different age/risk groups with different disease dynamics. We assume that the level of within-household transmission in a household of size n scales proportionally to $1/(n-1)$, though we acknowledge that true transmission is slightly more complex (Cauchemez et al., 2009).

Model. We consider independent households of size $n = 1, 2, \ldots, 8$, for each of which n_e stochastic simulations of the within-household epidemic are performed based on the Sellke construction (Andersson & Britton, 2000; Sellke, 1983). Given all infectious contacts outside lead to an actual infection because we are in the early phase of the epidemic and there is no depletion of susceptibles, each case infects, on average, R_g new cases outside. Inside the household, a case would infect on average R_h cases in an infinitely large household, but not all infectious contacts lead to real infections, given local saturation effects: in a household of size n, each infectious individual makes on average $R_h/(n-1)$ infectious contacts with each other specific individual throughout the infectious period, but only the first one will result in an infection, and only if the individual was susceptible at the time of contact.

Each individual is given an indicator function of whether they are symptomatic or not (individuals show symptoms independently of each other with probability p_s) and a resilience threshold. This last quantity is drawn from an exponential distribution with mean 1, and represents the overall infection pressure this individual is able to withstand before they get

Table 5
Information and values for each parameter in our differential equation based households model.

Parameter	Value	Details
Doubling time	3 days	Number of days until the number of cases doubles
$r_{p\to i}$	1/3	Inverse prodromal period
$r_{i\to\varnothing}$	1/4	Inverse infectious period
$r_{e\to p}$	1/5	Inverse latent period
H	0.8	Inverse exponential scaling of transmission with household size
α_W	0.65	Adherence to household isolation
E	0.65	Reduction in global transmission
τ_p	0.4	Secondary attack probability for a two-person household with one susceptible and one prodrome
τ_i	0.8	Secondary attack probability for a two-person household with one susceptible and one infective

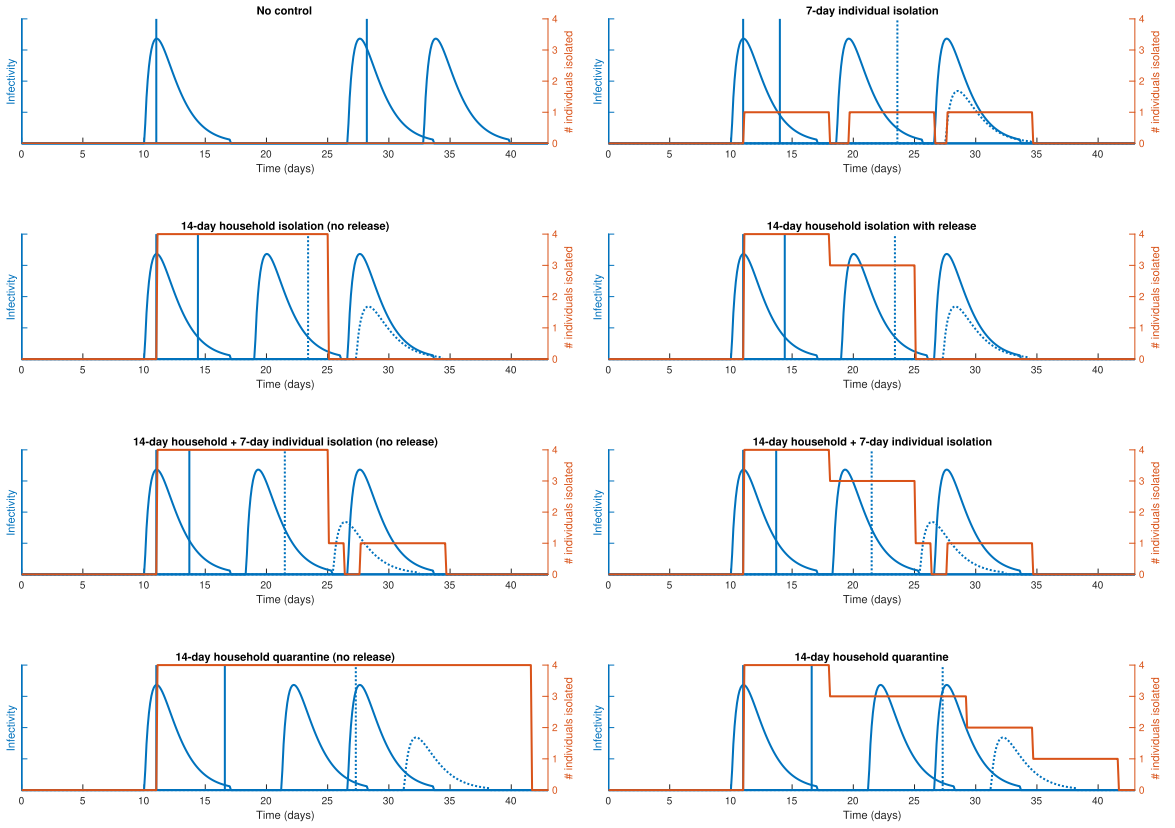


Fig. 8. Impact of various control policies on a single epidemic realisation in a household of size 4. The blue solid lines represent the infectivity of symptomatic individuals. Vertical solid lines represent the times of infection of symptomatic cases. Dashed curves and vertical lines represent the infectivity and time of infection of asymptomatic cases, which are assumed to be half as infectious as symptomatic ones. The total number of cases under isolation (possibly compounded, e.g. both household and individual isolation) is shown in red (right axis). All random numbers involved in the realisation of the stochastic epidemic are drawn at the start, before the impact of each control policy is implemented. Row 1 shows no isolation and individual and isolation. Rows 2, 3 and 4 show, respectively, household isolation, mixed isolation and household quarantine. The difference between the columns is that the basic policy on the left is "upgraded" to the more cost-effective version on the right that allows recovered individuals to leave the house as they cannot transmit outside anymore. When no control is implemented, the primary case (individual A, infected at time 0) infects another individual (B) around time 11. After a long latent period (i.e. incubation minus prodromal), B becomes infectious and infects a further individual (C). The last individual (D) escapes infection. When different intervention strategies are in place, within-household infectivity is increased. This can result in individual C becoming infected earlier in the outbreak and individual D no longer escaping infection, both due to the increased force of infection. In this simulation, the dynamics for individual B do not change since they are infected before A becomes symptomatic. Individual D is infected earliest under mixed isolation, because within-household transmission is higher than household isolation alone, due to increased adherence from individual isolation also being in place. Adherence levels to household quarantine are lower than those of household isolation, due to the higher demand of full quarantine, thus leading to less enhanced within-household transmission. We assume that adherence to individual isolation is 90%, household isolation is 80% and household quarantine is 60%. The more severe the intervention, the better it captures the infectious periods of infected individuals within the household.

infected. The infection pressure up to time τ is the integral from 0 to τ of the force-of-infection (FOI) applied to this individual coming from all infectious sources.

At the beginning of the within-household epidemic, a single initial case is assumed. Time is discretised with a predefined time step $dt = 0.1$ days. At any time step, the current infectivity of all infectives in that time step is summed over, keeping track differently of the infectivity spread outside and inside the household. An overall measure of the accumulated infectivity within the household is updated at each time step and when this crosses the resilience threshold of a susceptible individual, they acquire the infection.

We assume an individual spends half of their time outside and half inside the household. When self-isolation starts, the assumed adherence a_i represents the fraction of the time spent outside that is shifted from outside to within the household. Therefore, for perfect adherence, from the moment symptoms occur, the individual stops transmitting outside but their infectivity within the household grows by 100%. We also explore variations in this compensatory behaviour, so that the time of an individual is split in a more flexible proportion than $1:1$. The same argument applies to other control policies, with

Table 6
Information and values for each parameter in our non-Markovian households model.

Parameter	Value	Details
R_g	2.5	Basic reproductive ratio (outside household)
R_h	2.5	Basic reproductive ratio (within household)
R	0.245	Real-time growth rate
μ_E	4.84	Incubation period mean
σ_E	2.79	Incubation period standard deviation
d_p	1.5	Prodromal period duration
d_d	0.5	Delay from symptoms to isolation
d_i	7	Self isolation duration
μ_F	2.2	Infectivity mean
σ_F	1.64	Infectivity standard deviation
f_a	0.5	Relative infectivity of asymptomatic individuals

adherence levels a_h for household isolation and a_q for household quarantine. When multiple control policies are in place at the same time, their effect is assumed to be multiplicative: if an individual has symptoms and the household isolates, the outside transmission rate from that individual is reduced from baseline by a multiplicative factor $(1-a_i)\times(1-a_h)$ and the within-household transmission rate is the baseline value plus a fraction $1-((1-a_i)\times(1-a_h))$ of the baseline value. Therefore, implementing a control policy that reduces transmission outside might lead to more infections in the household (see Fig. 8 and the associated accumulated infection pressure in Fig. 10).

We denote by $\beta_g^n(\tau)$ the average global infectivity profile of a household of size n, i.e. the time-point average of the rates at which new cases outside are generated by any case infected in all simulated epidemics in a household of size n. During the exponentially growing phase, any global infection starts a new within-household epidemic. Furthermore, larger households are more likely to be infected because they have more members. Therefore, if h_n is the probability that a randomly selected household has size n,

$$\pi_n = \frac{nh_n}{\sum_m mh_m}$$

gives the probability that the household of a randomly selected individual is of size n. This is called the size-biased distribution. The global infectivity profile of the average household infected during the exponentially growing phase is then

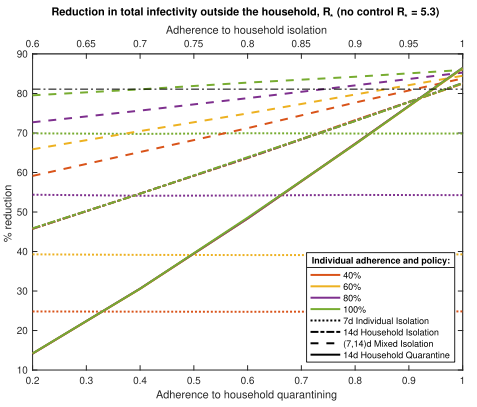
$$\beta_g(\tau) = \sum_n \pi_n \beta_g^n(\tau)$$

and the area under this curve is known in the literature as the household reproduction number, and is typically denoted by R_*. If enough transmission is prevented, so that $R_* < 1$, the epidemic is controlled. The basic reproduction number R_0 and R_* share the same threshold at one, so they are simultaneously larger, equal or smaller than unity. However, in a growing epidemic $R_0 < R_*$ (Ball et al., 2016; Pellis et al., 2012). The real-time growth rate r is related to R_* by the Lotka-Euler equation

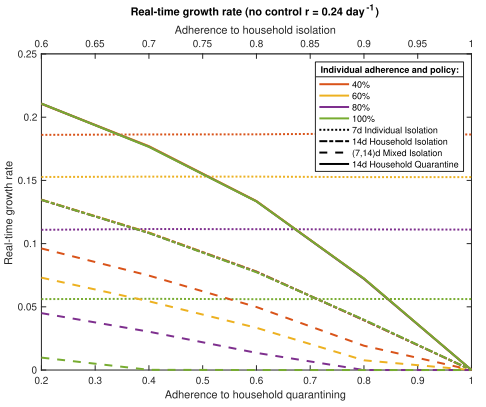
$$\frac{1}{R_*} = \int_0^\infty \beta_g(\tau) e^{-r\tau} d\tau.$$

Parameterisation. At baseline we take $R_g = 2.5$ and $R_h = 2.5$, which gives a real-time growth rate r of about 0.245. Each individual, irrespective of whether they will be infected or not, is given (independently of each other) a duration of: incubation period (randomly drawn from a Gamma distribution with mean $\mu_E = 4.84$ and standard deviation $\sigma_E = 2.79$ days), prodromal period ($d_p = 1.5$ days), and delay from onset of symptoms to when the individual detects the symptoms and enters isolation ($d_d = 0.5$ days), and the period of self-isolation ($d_i = 7$ days). At the end of the incubation period, a symptomatic individual starts showing symptoms, which allow the triggering of control policies (after the delay d_d). We assume that two thirds of individuals go on to develop symptoms, with the rest remaining asymptomatic. Asymptomatic individuals do not trigger any policy. After a latent period (defined as the incubation minus the prodromal period) and irrespective of symptoms, any infected case starts an infectious period with an infectivity that changes over time following the probability density function of a Gamma distribution (mean $\mu_F = 2.2$ and standard deviation $\sigma_F = 1.64$ days). Asymptomatic cases are assumed half as infectious as symptomatic ones (relative infectivity $f_a = 0.5$).

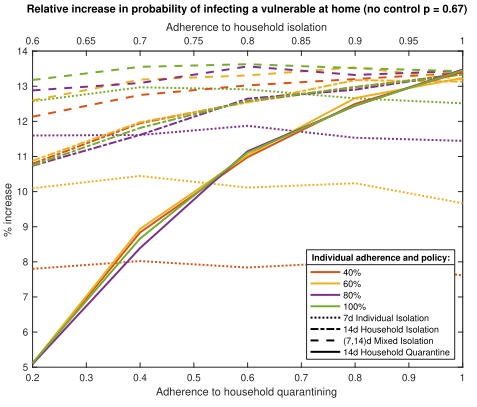
Summary. Under the baseline parameter values (Table 6), control can in principle be achieved via certain interventions, but only for high levels of adherence, which might be difficult to enforce for a prolonged length of time (Fig. 9a). More importantly, the model's conclusions are highly sensitive to variations in parameter choices, which are uncertain. Parameters that present problems here are the delay from symptom onset to isolation (with control failing for 1 day detection delay unless



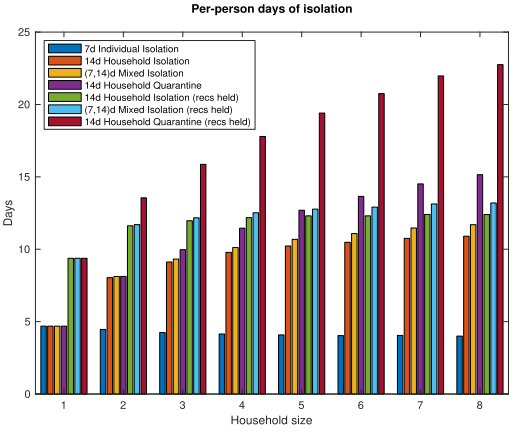
(a) Percentage reduction in R_*.



(b) Reduction in the real time growth rate.



(c) Increase in risk of infection an initially susceptible vulnerable person experiences in the household.



(d) Average number of days of isolation a person experiences in households of different sizes.

Fig. 9. Impact of different control policies and levels of adherence on transmission, infection risk, and time in isolation. (a) Percentage reduction in R_*, defined as the total amount of community transmission spread by an average household early in the epidemic, which equals 5.3 for baseline parameters in the absence of control; (b) real-time growth rate, which is assumed to be 0 (rather than negative) when the infection is controlled; (c) increase in risk of infection an initially susceptible vulnerable person experiences in the household; and (d) the average number of days of isolation a person experiences on average in households of different sizes, computed for each size as the average total person-days in isolation divided by the number of individuals in the household. In (a)–(c), line styles refer to different control policies and colours to different levels of adherence to individual isolation. The lower x-axis gives the adherence to household quarantine, and the upper x-axis adherence to household isolation. We assume that household isolation is less demanding, and therefore adherence is assumed to be "twice as high", meaning it is at the midpoint between that of household quarantine and 1 (e.g. 0.6 for an x value of 0.2, 0.9 for an x value of 0.8, etc.). The black dash-dotted line in (a) gives the amount needed to control the spread by achieving $R_* = 1$. Notice how: the effect of individual isolation is independent of adherence to household quarantine (dotted lines); the effect of household isolation is independent of adherence to individual isolation (overlapping dash-dotted lines); mixed isolation is always superior to household isolation; household quarantine is only optimal at really high levels of adherence (for these baseline parameters, generally, beyond the level needed to achieve control), but quickly becomes suboptimal to mixed isolation as adherence is reduced. When a sufficiently large reduction in R_* is achieved in (a), the growth rate drops to 0 in (b). The increased risk an initially susceptible vulnerable person is infected at home (c) does not reflect this effect, as it represents the increased risk conditional on an introduction: if the infection were controlled in the community, the overall risk of a vulnerable person getting infected would vanish as the risk of introduction in the household vanishes. For these plots, $n_e = 10000$ simulations are performed for each household size. Nevertheless, a large amount of stochastic noise is still visible in (c). In (d), the same control policies are considered, but the household-based ones are considered both in their naïve form (where recovered individuals remain isolated), and in their upgraded version where recovered individuals are free to leave the house: they are identical in terms of transmission but the naïve versions are significantly more costly in terms of person-days of isolation. In a household of size 1 (no within-household transmission), the days in isolation would be exactly 7 or 14 if all cases were symptomatic (here $p_s = 2/3$); similarly, in all households, individual isolation would total exactly 7 days if all cases were symptomatic and all individuals in the household were ultimately infected. In (d), we assume that adherence is 100% to each intervention.

adherence is essentially perfect), proportion of asymptomatic infections (any chance of control lost at 50%) and the strength of asymptomatic transmission. The short delay before symptomatic individuals isolate may be unrealistic unless the susceptible population is very well-informed about symptoms that call for isolation, and so likely does not apply in very early stages of an outbreak. Overall, in the face of the many uncertainties, household-based interventions triggered purely by symptoms appear useful to slow the spread but need to be complemented by other policies.

Comparing the different strategies (Fig. 9b), household quarantine can be optimal (as one might expect), but this requires high adherence levels. As adherence drops, this strategy becomes suboptimal to mixed isolation. Mixed isolation is

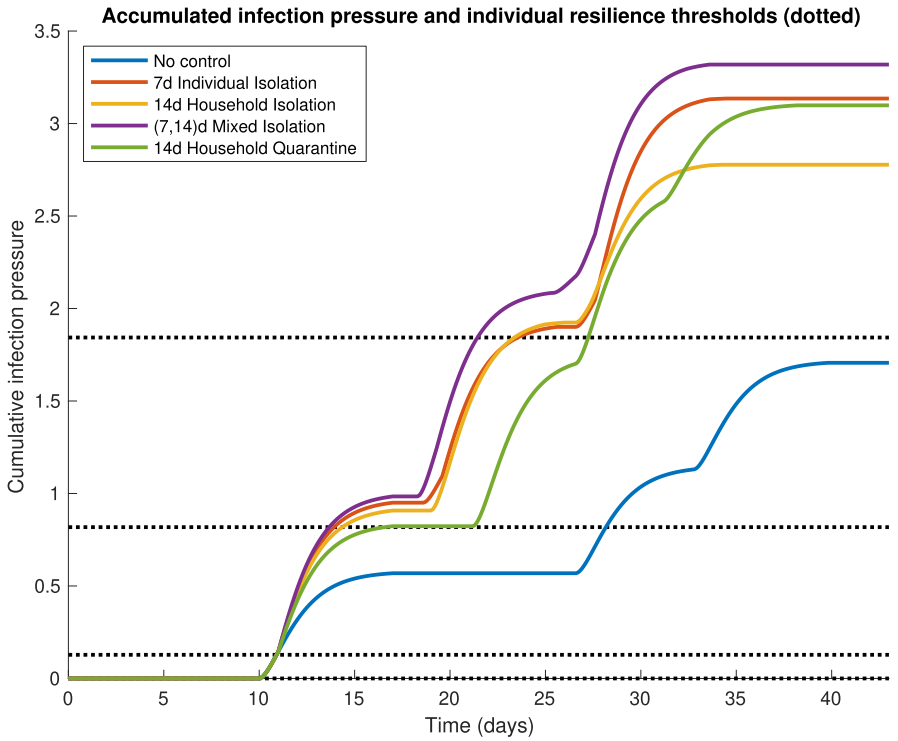


Fig. 10. Accumulated infection pressure in the simulation presented in Fig. 8 for different control policies. Horizontal dotted lines represent individuals' resilience thresholds. As time progresses, the accumulated infection pressures (coloured lines) increase and when they cross the resilience thresholds, the corresponding individual acquires infection. Notice that: in the absence of control, one individual escapes infection; with household isolation only, the infection pressure reaches a relatively low endpoint because of the last symptomatic individual slipping through and not transmitting much in the household; with mixed isolation, infection pressure is higher due to combined adherence; and with household quarantine, the infection pressure builds up more slowly at the beginning due to lower adherence. We assume that adherence to individual isolation is 90%, household isolation is 80% and household quarantine is 60%.

significantly better than household isolation on its own and requires little extra social cost, so should not cause adherence to drop (relative to household isolation adherence levels). The difference between the two strategies comes down to the transmission slipping through after the 14 day household isolation. The cheapest strategy, when considering working age adults, is individual isolation (Fig. 9d), but the effect is limited compared to the other models and cannot achieve control in the baseline scenario even with 100% adherence.

Overall, the mixed isolation strategy appears to be most cost-effective. However, this is dependent on the assumption that adherence is better for 14 day isolation rather than a very long quarantine. It can be observed that household-based interventions are more effective than individual isolations, demonstrating the importance of these strategies in designing intervention policy. Fig. 8 shows how the different isolation strategies contain the infectious periods of individuals within the household and also indicates the number of individuals being isolated within the household.

To study the impact such an increased within-household transmission has on the chance that a vulnerable individual is infected in the household, we randomly choose one non-primary case in the household as the vulnerable one and count how many of the n_e epidemics result in this individual being infected under the different control policies (Fig. 9c). Under these interventions, the risk of a vulnerable individual getting infected within-household, conditional on the infection entering it in the first place, is in the range $5-15\%$.

Since this model relies on the Sellke construction, we calculate the infection pressure that accumulates (within a household) during the outbreak. In relation to Fig. 8, we report in Fig. 10 the infection pressure that accumulates for the different control policies, showing the different impact each intervention can have on the within household dynamics.

3.4. Extinction probabilities

Social distancing, isolation and lockdowns act to mitigate the spread of an infectious disease and reduce the number of cases. However, such interventions, particularly widespread lockdowns, cannot be maintained indefinitely and must be lifted at some point. For the disease to be controlled, these interventions can be implemented until pharmaceutical interventions

are developed, such as a vaccine, or until the case numbers are low enough that the disease may go extinct. Here, we consider the situation where interventions are lifted just before extinction, when the number of cases has reached a low but non-zero initial value n_0: at this point, the number of cases might rebound or might go extinct by random chance despite an $R_0 > 1$. We use a time-inhomogeneous birth-death chain model (Kendallet al., 1948) to investigate the probability of extinction in this context. The n_0 "initial" cases give rise to new cases at a time-dependent rate $\beta(t)$ and recover at rate $\delta(t)$. Letting $Z(t)$ denote the random variable that gives the number of cases at time t, we are first interested in obtaining an expression for the probability generating function

$$Q(t,s) = \mathbb{E}\left[s^{Z(t)}\right] = \sum_{n=0}^{\infty} \mathbb{P}(Z(t)=n)s^n.$$

It follows that $Q(t,s)$ satisfies the differential equation

$$\frac{\partial Q}{\partial t} = \left(\delta(t) - (\delta(t)+\beta(t))s + \beta(t)s^2\right)\frac{\partial Q}{\partial s},$$

subject to the initial condition

$$Q(0,s) = s.$$

Solving for Q and setting $s = 0$ gives the probability that, at time t, the number of cases has reached zero and the disease has become extinct. We denote this probability by $q(t)$ (Alexander & Bonhoeffer, 2012), which is given by

$$q(t) = 1 - \left(\int_0^t \left(\beta(t_1)e^{I(t_1)}\right)dt_1 + e^{I(t)}\right)^{-1},$$

where $I(t) = \int_0^t \delta(t_1) - \beta(t_1)dt_1$.

The above case considers a closed population. Since the virus has spread worldwide, for any population of interest, immigration of infected individuals cannot be ignored. To capture this, we model the case where immigration from external sources is introduced into the system at a rate $\eta(t)$, and are similarly interested in the random variable $Y(t)$, which denotes the number of cases at time t. The corresponding generating function, $R(t,s)$, for this random variable satisfies

$$\frac{\partial R}{\partial t} = \left(\delta(t) - (\delta(t)+\beta(t))s + \beta(t)s^2\right)\frac{\partial R}{\partial s} + \eta(t)(s-1)R.$$

Again, solving for $R(t,s)$ and setting $s = 0$ gives the probability, $r(t)$, that there are no cases of infected individuals left, at which time a new case can only arise through immigration from an external source. This probability is given by

$$r\left(t\right) = \exp\left(-\left(1-q\left(t\right)\right)\int_0^t \eta(t_1)dt_1\right)$$

We simulate data based on one initial case $n_0 = 1$, though this may easily be extended to any number of initial cases. We run simulations both with and without immigration, choosing $\beta(t) = 3/(7(1+5e^{-t}))$ and $\delta(t) = 1/7$ for all t, so that an effective reproduction number given by $\beta(t)/\delta(t)$ grows gradually from 0.5 to 3 after interventions are released, and choosing immigration rate $\eta(t) = W_0e^{-t}$, where W_0 is the initial (constant) rate of importation of cases before any controls on immigration are put into effect. We set $W_0 = 5$ imported cases per day. With these choices of parameters, the resulting extinction probabilities are given in Fig. 11. Note that we are assuming the immigration rate is decreasing to 0, so if the infection is controlled internally for long enough, an overall ultimate extinction is possible in this model. For these parameter choices, the final probability of extinction, defined as $\lim_{t\to\infty} q(t)$ (without immigration) and $\lim_{t\to\infty} r(t)$ (with immigration) are approximately 0.446 and 0.002, respectively. It should be noted that $q(t)$ concerns the best case scenario with only one initial case. Increasing the number of initial cases n_0 scales the probability of extinction by $q(t)^{n_0}$. These probabilities suggest that, without widespread immunity, stochastic extinction might be aided by social distancing but is heavily compromised by immigration. Border controls, therefore, if of limited use when transmission is self-sustaining, become key when the number of cases is low. Note that we have assumed an importation function $\eta(t)$ that goes to 0 for large t, in line with a pandemic that goes extinct in other geographical regions. However, the presence of an animal reservoir might lead to an importation function that is non-zero over longer time scales, thus effectively making ultimate extinction impossible unless the effective reproduction number is kept below one by a systematic and permanent intervention (e.g. technology-based change in behaviour) or herd immunity.

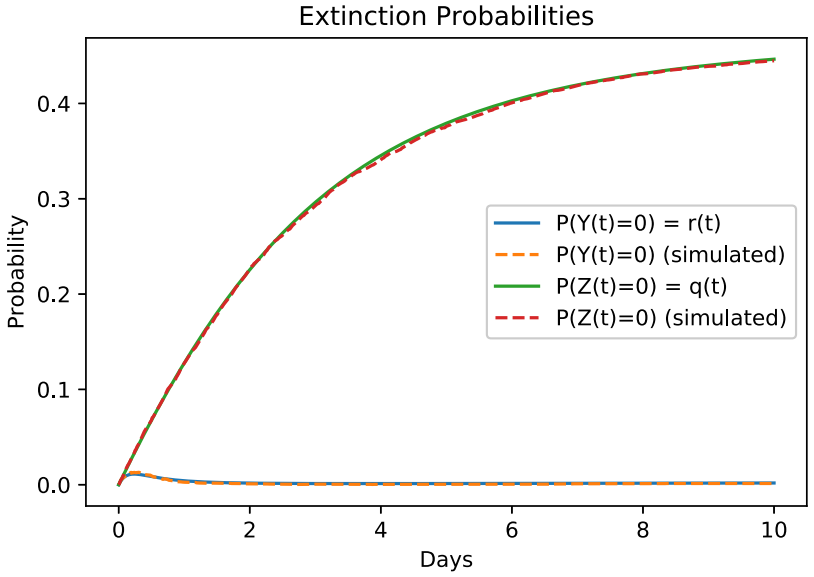


Fig. 11. Extinction probabilities, both analytic and simulated, for the choices of $\beta(t), \delta(t)$ and $\eta(t)$ described in the main text. The simulated extinction probabilities were calculated from 10,000 simulations of a birth-death chain both with and without immigration, for which the code can be obtained via the supplementary material.

3.5. Contact tracing and household isolation

Contact tracing is a complementary control policy to isolation or quarantine. When a case is discovered, attempts are made to identify and isolate individuals who may have been infected. In doing so, some of the secondary cases will be discovered and isolated early in their infection, decreasing their effective infectious period. If contact tracing is successful, it can greatly reduce the effective reproduction number of the infection, and in combination with other interventions may drive an epidemic extinct, as was seen in the case of SARS (Wilder-Smith, Chiew, & Lee, 2020).

Contact tracing in itself presents numerous challenges, which are exacerbated by its success relying not only on the effectiveness of the tracing process but also the underlying transmission characteristics. For COVID-19, some of these challenges include mild symptoms which cause infections not to be reported, pre-symptomatic transmission which occurs before a case is reported, and short generation times (Ganyani et al., 2020) which can cause the epidemic to outrun contact tracing. Additionally, contact tracing is only feasible for smaller case numbers, because each case generates multiple contacts to follow up, so the tracing workload expands dramatically, and an increasing number of chains remain unobserved. This makes it a viable strategy in the early days of an outbreak, or, if containment has failed, following a period of severe interventions, such as a lockdown. Combining contact tracing with isolation is being considered by many countries as part of a test, trace and isolate strategy to be implemented once lockdowns or comparable measures are lifted, provided these lockdowns succeed at driving case numbers sufficiently low. In this section, we develop a household-level contact tracing model for an emerging outbreak, since we do not wish to make assumptions about immunity or depletion of susceptibles. These assumptions can be added to the model as the availability of data into immunity improves. We are interested in the likelihood that the contact tracing process is overwhelmed by large case numbers and the likelihood that, combined with isolation, it can drive the disease to extinction.

The early days of an outbreak can be modelled using a branching process, where generations of infections produce infectious offspring. Contact tracing processes can be incorporated as a superinfection along the tree generated by the branching process (Ball, Knock and O'Neill, 2015). When a node is 'superinfected' by the contact tracing process, it is isolated.

We model the infection spreading through a fully susceptible population of individuals, segmented into households of different sizes according to the 2019 ONS survey (Office for National Statistics, 2019), and progress through discrete time steps of 1 day. As such, our branching process is at the household level, coupled with localised within-household epidemics. This allows us to model contact tracing strategies that isolate whole households, which may contain several undetected infections. It also enables a wider range of contact tracing strategies to be modelled, each with different intervention scope and costs.

Each day, individuals (or nodes) make contacts to a random set of individuals; divided into local contacts to members of the same household, and global contacts to members of other households. The number of individuals contacted in a day is distributed using an overdispersed negative binomial distribution and parameterised using estimates from the POLYMOD social contact survey (Mossong et al., 2008), stratified by household size. Since the probability that a contact causes infection

cannot be directly observed, we use improper hazard rates that give rise to the 5 day COVID-19 generation time (Ferretti et al., 2020) and $R_0 = 3$.

For contact tracing to begin, an infection must be diagnosed, which we assume occurs 70% of the time among infected individuals due to flaws in reporting or very mild symptoms in those infected. We assume a Gamma distributed incubation period with mean 4.84 (Table 2) and a Geometric reporting delay from symptom onset with mean 4.8 days (Kraemer et al., 2020). Intuition suggests that if $R_0 = 3$ then tracing two thirds of contacts will control the epidemic. However, in practice transmission may occur before tracing, so this will not reduce the number of infectious contacts by two thirds. To demonstrate this, we assume that contact tracing successfully traces two thirds of contacts. Trained professionals have to trace all reported contacts from the last 14 days, so we assume that the contact tracing delay follows a Geometric distribution with a mean of 2 days. Individuals are considered recovered 21 days after infection, as the chances that they are still transmitting then are negligible.

Though our general framework can be modified extensively, we assume the following contact tracing strategy. When an individual reports infection, their household is immediately isolated. Contact tracing attempts are then made for all households connected to one of the individuals in this household, whether symptomatic or not. When a connected household is identified (after the contact tracing delay), all individuals within the household are immediately placed under observation. If any of the individuals in the observed households develop symptoms, then the household becomes isolated and the contact tracing process continues to connected households. When a household is isolated, we assume all individuals are isolated with 100% adherence, and cannot transmit the virus within or outside the household. The assumption that isolation prevents local infections is unrealistic, but does not change the overall behaviour of the process as there are no more global infections. This strategy imposes high individual-level cost, since by isolating all individuals within a household, it isolates individuals who have not had direct contact with an infected individual. In practice, such a strategy may have poor adherence. Fig. 13a shows an example contact tracing network.

3.5.1. Hitting times of contact tracing capacities

When choosing contact tracing strategies, a balance must be struck between the effectiveness of a strategy and the resources that it requires. Some strategies are only feasible when there are few infections, since the resources required can grow rapidly depending on the dynamics of the outbreak and the contact tracing process.

To define the capacity of the contact tracing process, we consider the ability of a public health agency to observe the condition of those asked to self-isolate, due to their recent exposure to an infected individual. The health agency must remain in contact for the duration of the 14 day self-isolation period, so that if any individual under isolation develops symptoms and then tests positive, the contact tracing process can be initiated on this node. We will define the capacity of the contact tracing process to be the number of people that can be placed under observation and assume two possible capacities: 800 and 8000. We assume that when a node is contact traced, they are asked to report their global contacts for the last 14 days. All global contacts are assumed to be to a new person since we are in the early stages of an outbreak. Parameters are given in Table 8.

We carried out 6507 simulations of the contact tracing process for 150 days. Contact tracing capacity was reached in 5000 simulations, and in 180 the epidemic neither went extinct nor was the 8000 capacity reached. In the remaining simulations, the epidemic went extinct. Fig. 12a and Table 8 show that increasing the contact tracing capacity tenfold less than doubles the time until that capacity is reached. However, it does increase the odds of driving the epidemic to extinction without hitting the capacity by about 10% (Table 7). Different contact tracing strategies will strain different aspects of the health agency. A strategy that generates large amounts of work is only feasible if there are few active infections. The optimal strategy will need to compromise and may need to change depending on the number of active infections, which cannot be directly observed.

3.5.2. Extinction time

When there is a small number of cases in a single country, it may be possible to drive the pathogen to extinction. This small case number could correspond to the start of an outbreak or removing of severe interventions. We consider the latter case, but conservatively assume a fully susceptible population.

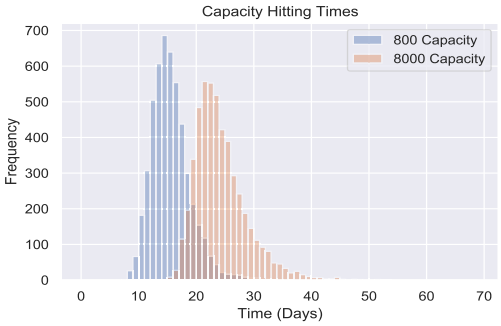
We assume that social distancing is enforced on day 0 and reduces global contacts by 70%. Full parameters are given in Table 8. Since we are interested in extinction, we will no longer consider the contact tracing capacity. Under these baseline parameter assumptions and 10,000 simulations, the combined force of this contact tracing strategy and isolation is enough to drive the epidemic extinct (Fig. 13b), but measures will need to be in place for months in some cases. If the infection is ever re-imported, then the process would begin again, since herd immunity is not achieved. Note that the minimum extinction time is 21 days due to this being the time after which an infected individual is labelled recovered.

Table 7
Contact tracing capacity hitting probability and hitting time distribution.

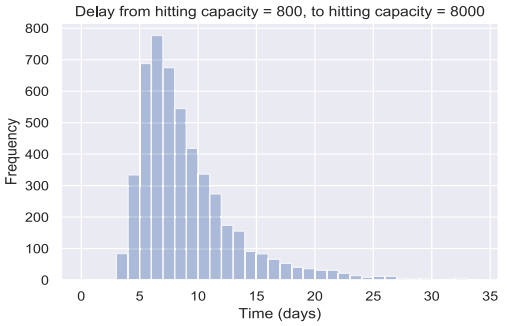
Quantity	Results
Mean (time 800 capacity reached)	13.9 days
Mean (time 8000 capacity reached)	22.5 days
Hitting probability (800)	81.2%
Hitting probability (8000)	76.8%

Table 8
Contact tracing model - parameter table.

Parameter	Value	Details
R_0	3	Basic reproductive ratio
Generation time	5 days	Mean time from being infected to infecting another individual
Incubation period	4.84 days	Mean time from being infected to developing symptoms
Reporting delay	4.8 days	Mean time from developing symptoms to reporting to a healthcare system
Diagnoses rate	70%	Proportion of cases that are successfully diagnosed
Contact tracing success rate	66.67%	Proportion of contacts that are successfully traced
Contact tracing delay	2 days	Mean length of time taken to trace contacts after reporting of a case
Time to recovery	21 days	Length of time until individuals are taken to no longer be infectious
Global reduction	70%	Reduction in transmission caused by large-scale global interventions

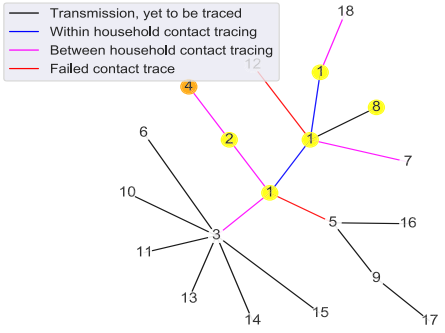


(a) Distribution of hitting times for 800 (blue) and 8000 (orange) contacts traced. The distributions appear to follow Gumbel distributions.

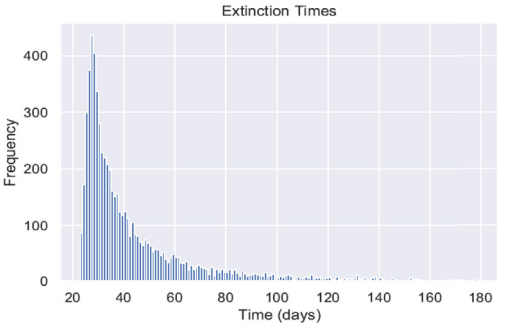


(b) The delay between hitting 800 contacts traced and 8000. This distribution appears to follow a Gumbel distribution.

Fig. 12. Capacity hitting times for the contact tracing model.



(a) A visualization of the branching process model after 20 days. Node numbers represent the household that they belong to. Nodes 3, 7 and 18 will be contact traced but the process has not reached them yet, due to contact tracing delays. Node 4 has been contact traced, but has not been quarantined in the plotted scenario, as it does not display symptoms.



(b) Distribution of extinction times for epidemics that transmit the infection at least once. The distribution takes the shape of a generalised extreme value distribution, and there are several epidemics that still have not gone extinct after interventions have been in effect for long periods of time, which exemplifies how difficult it can be to make an epidemic extinct.

Fig. 13. Example of the contact tracing process (a) and the extinction times distribution (b) for the contact tracing model.

Additionally, this model only considers extinction under the assumption that no cases are imported. In Section 3.4, we have shown that importation of cases significantly reduces the extinction probability. This suggests that extinction may no longer be guaranteed, and the time to extinction will be significantly increased. This analysis has focused on a single contact tracing strategy using indicative parameters for COVID-19. The proposed model can be extended to more strategies and region specific parameters to inform the design of control policies. Also, as is shown in Table 8, contact tracing capacity is likely to be reached, which may prevent extinction from being achieved. This complication is compounded by the issues of loss of immunity or the presence of an animal reservoir discussed in Section 3.4.

4. Discussion

In this manuscript we have presented a range of mathematical tools to tackle infectious disease outbreaks. In particular, these tools address various technical questions posed by the authors to support the ongoing public health response to COVID-19. This toolkit considers both estimation efforts for key parameters, and investigative efforts (often numerical simulations) in gauging the effectiveness of various intervention or control measures. Joint consideration of estimation and simulation efforts is critical. Parameter estimates are obtained using a certain set of assumptions regarding the data, and investigations or simulations utilising these estimates should ensure that their underlying assumptions are consistent. These challenges in model construction and applicability of statistical methods are compounded by the limitations of the data with which decisions must be made. Some of the biases present in the data can be addressed with an improved data collection methodology - often challenging in the context of a fast-moving outbreak - but many are also inherent to the nature of early outbreak data (Britton & Scalia Tomba, 2019; Lesko, Keil, & Edwards, 2020 Edwards). The consequent lack of intuitive insight from this data underscores the need for careful parametric estimates, especially considering the large variability in predicted outcomes resulting from small differences in parameters. Even with robust estimates for some parameters, many other parameters are challenging to estimate using the available data. Therefore, models need to address this variability and uncertainty in order to inform public health policy.

We have presented methods to address biases arising from a growing force of infection, changes in the reporting rate, truncated data samples and a varying travel rate. We use these methods to account for these biases when estimating delay distributions, such as the incubation period, and the growth rate/doubling time. These biases can have significant impact when estimating key parameters: the mean incubation period estimates for COVID-19 range from 3.48 days without correcting for truncation to 4.84 days with the correction, and the doubling time in Hubei province decreases from 3.15 days without correcting for travel to 2.77 days. These differences can significantly alter our understanding of the outbreak, and could have a large impact on policy and public health. For instance, underestimating the incubation period may lead to quarantine strategies failing to identify infected individuals if the quarantine length is too short. Overestimating the doubling time (or underestimating the growth rate) will underestimate the risk posed to the host population - both in terms of final size of the epidemic and the rate at which it spreads, which can have significant public health impacts as discussed in (Pellis, Cauchemez, Ferguson and Fraser, 2020).

It is important to note that the above-mentioned biases, and consequent impact of implementing the methods correcting for their presence, may vary across different settings. As an example, the potential underestimation of the COVID-19 growth rate is exacerbated by an overlap in early outbreaks with a period of significant travel and movement in China, and would be less detrimental if first observed in other populations such as Italy. Also, for the incubation period, we have shown two different types of data; one from Wuhan and one from discrete infection events. In the Wuhan data set, truncation and force of infection biases are very important, whereas in the other data set, there is no force of infection bias since the infection events are observed.

When an outbreak occurs in an enclosed group, such as a large gathering, we may wish to know how many individuals are likely to be infected. We developed a statistical method to estimate the first generation size based on the number of symptomatic individuals, taking care to account for the uncertainty in this quantity. This ready reckoner can inform testing of large groups to help control the disease spread, but does not apply to later generations or the possible interventions enacted on the population.

Building on these enclosed population scenarios, we have developed a set of models that investigate public control measures or interventions on enclosed populations, such as households and care homes. These structured descriptions improve the population risk profiles relative to assumptions of homogeneous mixing. A complementary aspect to a structured population when modelling interventions is adherence. Motivated by vaccination modelling, we consider leaky adherence, where every household chooses to adhere or not whenever an event occurs, and all-or-nothing adherence, where some households adhere every time and some never adhere. We observed that in a homogeneous population, although the two types of adherence predict the same growth rate and final size, the timing of the peak and the early growth can be faster under all-or-nothing adherence. This insight, combined with lessons from the vaccination literature, suggests that efforts should focus on ensuring complete adherence in individuals or households with some level of pre-existing adherence, rather than pushing non-adherent individuals or households to change behaviour.

Dedicated modelling of disease spread in care homes is essential due to the documented history of co-morbidity of their residents during pandemics (Guan et al., 2020), (Guan et al., 2020; Wilder-Smith, Chiew, & Lee, 2020; Wu et al., 2020; Wu & McGoogan, 2020; Yang et al., 2020; Zhou et al., 2020). We do so by regarding care homes as closed populations that are subjected to a force of infection from an external epidemic. We develop a tool for analysing the risk posed to this population by determining the peak size of the epidemic within the care homes and the number of deaths. Applying this model to COVID-19, we find that by "cocooning" the care homes, i.e. shielding them to reduce the chance of introduction from the external outbreak, we can significantly reduce the size of the peak and therefore reduce the number of deaths. However, assessing the necessary level of shielding requires accurate characterisation of the external force of infection, and underestimating this may invalidate shielding efforts. A limitation to the proposed model is the deterministic within care home epidemic. However, since the average size of care homes is relatively large and we assume a high R_0 within care homes, the deterministic assumption is unlikely to significantly alter the conclusions.

When modelling households however, we are concerned with much smaller population sizes. Therefore, it is important to consider stochastic effects within each household, combined with between-household dynamics. We consider two different household models: one which contains features of both within- and between-household transmission, where small-scale transmission can be linked to the epidemic on a population level, and another which facilitates more detail in the within-household transmission and delay distributions, but with reduced correspondence to the population-wide transmission. With the first model, a 65% adherence to household isolation appears insufficient to control the epidemic without severe global reductions in transmission. Coupled with a short term global reduction, the epidemic can be controlled, but upon lifting the global intervention, which could take the form of a lockdown, household isolation is insufficient to maintain control. For the second model, we look into changing the strength of adherence, and the impact this can have on achieving control. Indicative but reasonable parameter values suggest that the COVID-19 outbreak can potentially be controlled using household isolation strategies, provided the level of adherence is sufficiently high. However, such a high level of adherence may be difficult to maintain in the long-term and this modicum of control is anyway highly sensitive to the chosen parameters. We further investigated the efficacy of various isolation or quarantine measures. A policy of individual isolation struggles to curtail the epidemic for any adherence. Instead, mixed isolation, whereby first the whole household isolates and any individual infected during isolation goes on to self isolate after household isolation is lifted, appears to be the most cost-effective strategy.

Countries have put into place strict social distancing and lockdown intervention to suppress or regain control of epidemics that threaten to overwhelm the health system and cause massive mortality, but they cannot be sustained in the long term without growing social and economic costs. We have shown however, that the probability of the epidemic becoming extinct once these policies are lifted, even when very few cases remain, is very small. We therefore consider a contact tracing intervention as a potential strategy for managing the COVID-19 outbreak, once severe lockdown interventions are lifted. We developed a household-level contact tracing model to explore the feasibility of combining these strategies to control the epidemic. Firstly, we noted that by using knowledge of household structure, we can reduce the burden on the contact tracing process by isolating household and removing them from the contact tracing process once an infected member has been identified. Secondly, we investigated how contact tracing combined with household isolation may drive the disease to extinction, finding that aggressive contact tracing coupled with household isolation can drive the epidemic to extinction under the indicative parameters assumed, when starting with a single infection. However, the time until extinction can be impractically long, which risks the contact tracing capacity being overwhelmed, suggesting such a strategy may be infeasible in practice. A less aggressive strategy could be implemented that would be less likely to overwhelm local health agencies. Whilst it may not lead to extinction, this can still be beneficial at mitigating and controlling the spread of an outbreak as part of a test, trace and isolate strategy.

There are many complexities when modelling an outbreak of a novel infectious disease. To address some of these, we have described a variety of techniques to serve as part of a generally applicable toolkit. However, our proposed models, and many other models, are subject to important limitations which must be considered prior to their application. Key among these is the lack of heterogeneous population mixing, such as through age-stratification (Pellis et al., 2020a) and different risk-groups (Valdano, Poletto, Boelle, & Colizza, 2019, pp. 1–18), and spatio-temporal variations (Lau et al., 2017), all of which influence modelling estimates and predictions. Nevertheless, the relative simplicity of the presented models allows for the development of qualitative intuition regarding the efficacy of various intervention methods, whilst providing tractable theoretical frameworks which can be further developed and better inform policy-makers.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

PD and TF are supported by the NIHR Manchester Biomedical Research Centre. LP, HS and CO are funded by the Wellcome Trust and the Royal Society (grant 202562/Z/16/Z). EF is funded by the MRC (grant MR/S020462/1). MF supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. TH is supported by the Royal Society (Grant Number INF/R2/180067) and Alan Turing Institute for Data Science and Artificial Intelligence. IH is supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Emergency Preparedness and Response and the National Institute for Health Research Policy Research Programme in Operational Research (OPERA) and Alan Turing Institute for Data Science and Artificial Intelligence.

Appendix A. Independence of the incubation period likelihood function and the reporting rate

To estimate the incubation period distribution, we need to find the distribution that maximises the probability of observing the sampled data. However, the sampled data does not directly record incubation period, and instead contains infection exposure window and the symptom onset date. Additionally, the sample does not contain all individuals, and therefore there is a reporting rate that must be incorporated into the likelihood function. If the reporting rate is constant, it can be ignored. However, in the data coming out the Wuhan, the reporting rate varies significantly, since individuals are no

longer exported from Wuhan after travel restrictions. Since the reporting rate depends on individuals leaving Wuhan, the main factor effecting the probability that a case is included in the data set is the date an individual leaves Wuhan. Therefore, the reporting rate depends on the days an individual spends in Wuhan and is independent of their symptom onset date. We need the likelihood function that an individual was infected between days a and b, had symptom onset on day y, and was included in the data set.

We will condition against the infection window, $a < I < b$, the case being included in the data set, $x \in D$, and symptom onset occurring before the truncation date, $Y < T$, and determine the probability that for such an individual we observe the given symptom onset date, $Y = y$. That is, we need to find $P(Y = y|\{a < I < b\} \cap \{Y \leq T\} \cap \{x \in D\})$. This can be rearranged to

$$\begin{aligned}
&\frac{P(\{Y=y\}\cap\{a<I<b\}\cap\{Y\leq T\}\cap\{x\in D\})}{P(\{a<I<b\}\cap\{Y\leq T\}\cap\{x\in D\})} = \frac{P(\{Y=y\}\cap\{x\in D\}|\{a<I<b\})}{P(\{Y\leq T\}\cap\{x\in D\}|\{a<I<b\})} \\
&= \frac{P(\{Y=y\}|\{a<I<b\})P(\{x\in D\}|\{a<I<b\})}{P(\{Y\leq T\}|\{a<I<b\})P(\{x\in D\}|\{a<I<b\})} \\
&= \frac{P(\{Y=y\}|\{a<I<b\})}{P(\{Y\leq T\}|\{a<I<b\})} \\
&= \frac{P(\{Y=y\}\cap\{a<I<b\})}{P(\{Y\leq T\}\cap\{a<I<b\})} \\
&= \frac{\int_a^b P(\{Y=y\}\cap\{I=i\})di}{\int_a^b P(\{Y\leq T\}\cap\{I=i\})di} = \frac{\int_a^b g(i) f_\theta(y-i)di}{\int_a^b g(i)\int_0^{T-i} f_\theta(x)dxdi},
\end{aligned}$$

where f_θ is the probability density function of the incubation period distribution. Therefore, the likelihood function $P(Y = y|a < I < b, Y \leq T)$ is independent of the reporting rate for the data coming out of Wuhan in the early days of the outbreak.

Appendix B. Generation-size derivation

We wish to determine the probability that the first generation has e_0 individuals, $E_0 = e_0$, given that i_τ symptomatic individuals are observed on day τ, $I_\tau = i_\tau$, which is given by

$$\begin{aligned}
&\mathbb{P}\left(E_0 = e_o | I_\tau = i_\tau\right) = \int_0^1 \mathbb{P}(E_0 = e_0 \cap P = p | I_\tau = i_\tau \Big) \mathrm{d}p \\
&= \int_0^1 \mathbb{P}(E_o = e_0 | I_\tau = i_\tau \cap P = p) \times \mathbb{P}(P = p | I_\tau = i_\tau) \mathrm{d}p \\
&= \int_0^1 \frac{\mathbb{P}(E_0 = e_0 \cap I_\tau = i_\tau | P = p)}{\mathbb{P}(I_\tau = i_\tau | P = p)} \times \mathbb{P}(P = p | I_\tau = i_\tau) \mathrm{d}p
\end{aligned}$$

To solve this, we need to determine the distribution of the infection probability P given the number of observed symptomatics. Assuming that P is uniformly distributed, we have

$$\begin{aligned}
&\mathbb{P}(P = p | I_\tau = i_\tau) \propto \mathbb{P}(I_\tau = i_\tau | P = p) \\
&= \binom{n}{i_\tau} (pF(\tau))^{i_\tau} (1 - pF(\tau))^{n - i_\tau} \\
&\mathbb{P}(P = p | I_\tau = i_\tau) = c \times \binom{n}{i_\tau} (pF(\tau))^{i_\tau} (1 - pF(\tau))^{n - i_\tau},
\end{aligned}$$

where $F(\cdot)$ is the cumulative density function for the incubation period, n is the number of initially exposed individuals, and c is a normalising constant such that

$$c = \left(\int_0^1 \binom{n}{i_\tau} (pF(\tau))^{i_\tau} (1 - pF(\tau))^{n-i_\tau} \mathrm{d}p \right)^{-1}$$

$$= \frac{\left(\binom{n}{i_\tau} \right)^{-1} (i_\tau + 1)(F(\tau))^{-i_\tau}}{{}_2F_1(i_\tau + 1, i_\tau - n, i_\tau + 2, F(\tau))}$$

$$\mathbb{P}(P = p | I_\tau = i_\tau) = \frac{p^{i_\tau}(i_\tau + 1)(1 - pF(\tau))^{n-i_\tau}}{{}_2F_1(i_\tau + 1, i_\tau - n, i_\tau + 2, F(\tau))},$$

where ${}_2F_1$ represents the hypergeometric function (Abramowitz & Stegun, 1948). Substituting this into Equation (3) gives

$$\mathbb{P}(E_0 = e_o | I_\tau = i_\tau) = \int_0^1 \frac{\binom{n}{e_0} p^{e_0} (1-p)^{n-e_0} \binom{e_0}{i_\tau} F(\tau)^{i_\tau} (1 - F(\tau))^{e_0 - i_\tau}}{\binom{n}{i_\tau} (pF(\tau))^{i_\tau} (1 - pF(\tau))^{n-i_\tau}}$$

$$\times \frac{p^{i_\tau}(i_\tau + 1)(1 - pF(\tau))^{n-i_\tau}}{{}_2F_1(i_\tau + 1, i_\tau - n, i_\tau + 2, F(\tau))} \mathrm{d}p$$

$$= \int_0^1 p^{e_0} (1-p)^n \times \left(\frac{1 - F(\tau)}{1 - p} \right)^{e_0} \times (1 - F(\tau))^{-i_t} \frac{(n - i_\tau)!(i_\tau + 1)}{(n - e_0)!(e_0 - i_\tau)!}$$

$$\times \frac{1}{{}_2F_1(i_\tau + 1, i_\tau - n, i_\tau + 2, F(\tau))} \mathrm{d}p,$$

which simplifies to

$$\mathbb{P}(E_0 = e_o | I_\tau = i_\tau) = \frac{e_0!}{(e_0 - i_\tau)!} (1 - F(\tau))^{e_0} \times \frac{(1 - F(\tau))^{-i_\tau} (n - i_\tau)!(i_\tau + 1)}{(n+1)! {}_2F_1(i_\tau + 1, i_\tau - n, i_\tau + 2, F(\tau))}.$$

This gives a distribution of the generation-size based on the number of observed symptomatic individuals by time τ.

Appendix C. Estimating the generation size using a lower bound on the number of symptomatic individuals

In the analysis in Section 2.4, it is assumed that every person who has developed symptoms by time τ is known to the observer. However, depending on the disease, symptoms can be subjective. One person may not notice something another person may visit hospital for. Additionally, one person may not want to come forward with symptoms if they are worried about the repercussions of coming forward (for example being isolated against their own will).

To address this, rather than considering I_τ as the total number of people who have developed symptoms by time τ, we can consider $\tilde{I}_\tau$ as the number of people who have presented with symptoms by time τ. We do no know the true value of I_τ, but we know that it cannot be below $\tilde{I}_\tau$. We assume that the probability of $\tilde{I}_\tau$ being $\tilde{i}_\tau$ for a given value of i_τ is uniform at $\frac{1}{i_\tau + 1}$. We can then use the same methods as above to infer a distribution for P.

$$\mathbb{P}(\tilde{I}_\tau = \tilde{i}_\tau | I_\tau = i_\tau) = \begin{cases} \frac{1}{i_\tau + 1}, & \text{if } 0 \le \tilde{i}_\tau \le i_\tau \\ 0, & \text{otherwise} \end{cases}$$

C.1. Inferring a distribution for P given $\tilde{I}_\tau = \tilde{i}_\tau$

We again assume an uninformative uniform prior distribution for P and as a result the posterior distribution for P is proportional to the likelihood that $\tilde{I}_\tau = \tilde{i}_\tau$ given $P = p$:

$$\mathbb{P}(P = p | \tilde{I}_\tau = \tilde{i}_\tau) \propto \mathbb{P}(\tilde{I}_\tau = \tilde{i}_\tau | P = p) = \sum_{i_\tau = \tilde{i}_\tau}^{n} \mathbb{P}(I_\tau = i_\tau | P = p) \times \frac{1}{i_\tau + 1}$$

$$= \sum_{i_\tau=\tilde{i}_\tau}^{n} \binom{n}{i_\tau} (pF(\tau))^{i_\tau} (1-pF(\tau))^{n-i_\tau} \frac{1}{i_\tau+1} = \binom{n}{\tilde{i}_\tau} \frac{(pF(\tau))^{\tilde{i}_\tau}}{\tilde{i}_\tau+1} (1-pF(\tau))^{n-\tilde{i}_\tau} {}_2F_1\left(1, \tilde{i}_\tau - n, \tilde{i}_\tau + 2, \frac{-pF(\tau)}{1-pF(\tau)}\right)$$

$$\mathbb{P}(P=p|\tilde{I}_\tau = \tilde{i}_\tau) = c \times$$

$$\binom{n}{\tilde{i}_\tau} \frac{(pF(\tau))^{\tilde{i}_\tau}}{\tilde{i}_\tau+1} (1-pF(\tau))^{n-\tilde{i}_\tau} {}_2F_1\left(1, \tilde{i}_\tau - n, \tilde{i}_\tau + 2, \frac{-pF(\tau)}{1-pF(\tau)}\right)$$

where c is a normalising constant such that

$$c = \left(\int_0^1 \binom{n}{\tilde{i}_\tau} \frac{(pF(\tau))^{\tilde{i}_\tau}}{\tilde{i}_\tau+1} (1-pF(\tau))^{n-\tilde{i}_\tau} {}_2F_1\left(1, i_l - n, i_l + 2, \frac{-pF(\tau)}{1-pF(\tau)}\right) dp \right)^{-1}$$

$$= \left(\binom{n}{i_l} \frac{F(\tau)^{\tilde{i}_\tau}}{\tilde{i}_\tau+1} \int_0^1 p^{\tilde{i}_\tau} (1-pF(\tau))^{n-\tilde{i}_\tau} {}_2F_1\left(1, \tilde{i}_\tau - n, \tilde{i}_\tau + 2, \frac{-pF(\tau)}{1-pF(\tau)}\right) dp \right)^{-1}$$

$$\mathbb{P}P = p \left| \tilde{I}_\tau = \tilde{i}_\tau \right) = p^{\tilde{i}_\tau} 1 - pF(\tau))^{n-\tilde{i}_\tau} {}_2F_1\left(1, \tilde{i}_\tau - n, \tilde{i}_\tau + 2, \frac{-pF(\tau)}{1-pF(\tau)}\right)$$

$$\times \left(\int_0^1 y^{\tilde{i}_\tau} \left(1-yF(\tau)\right)^{n-\tilde{i}_\tau} {}_2F_1\left(1, \tilde{i}_\tau - n, \tilde{i}_\tau + 2, \frac{-yF(\tau)}{1-yF(\tau)}\right) dy \right)^{-1}$$

However, this integral cannot be found analytically and instead must be calculated numerically.

C.2 Inferring a distribution for E_0 given $\tilde{I}_\tau = \tilde{i}_\tau$

We can use the same probability analysis in Section 2.4 to write the following probability formula:

$$\mathbb{P}\left(E_0 = e_0 \left| \tilde{I}_\tau = \tilde{i}_\tau \right.\right) = \int_0^1 \frac{\mathbb{P}(E_0 = e_0 \cap \tilde{I}_\tau = \tilde{i}_\tau | P=p)}{\mathbb{P}(\tilde{I}_\tau = \tilde{i}_\tau | P=p)} \times \mathbb{P}\left(P=p \left| \tilde{I}_\tau = \tilde{i}_\tau \right.\right) dp$$

$$= \int_0^1 \frac{\binom{n}{e_0} p^{e_0} (1-p)^{n-e_0} \sum_{i=\tilde{i}_\tau}^{e_0} \binom{e_0}{i} F(\tau)^i (1-F(\tau))^{e_0-i} \times \frac{1}{i+1}}{\binom{n}{\tilde{i}_\tau} \frac{(pF(\tau))^{\tilde{i}_\tau}}{\tilde{i}_\tau+1} (1-pF(\tau))^{n-\tilde{i}_\tau} {}_2F_1\left(1, \tilde{i}_\tau - n, \tilde{i}_\tau + 2, \frac{-pF(\tau)}{1-pF(\tau)}\right)}$$

$$\times p^{\tilde{i}_\tau} \left(1-pF(\tau)\right)^{n-\tilde{i}_\tau} {}_2F_1\left(1, \tilde{i}_\tau - n, \tilde{i}_\tau + 2, \frac{-pF(\tau)}{1-pF(\tau)}\right)$$

$$\times \left(\int_0^1 y^{\tilde{i}_\tau} \left(1-yF(\tau)\right)^{n-\tilde{i}_\tau} {}_2F_1\left(1, \tilde{i}_\tau - n, \tilde{i}_\tau + 2, \frac{-yF(\tau)}{1-yF(\tau)}\right) dy \right)^{-1} dp$$

$$= \int_0^1 \frac{\left(\frac{p(1-F(\tau))}{1-p}\right)^{e_0} \left(\frac{1}{1-F(\tau)}\right)^{\tilde{i}_\tau} (1-p)^n (n-\tilde{i}_\tau)! {}_2F_1\left(1, \tilde{i}_\tau - e_0, \tilde{i}_\tau + 2, \frac{-F(\tau)}{1-F(\tau)}\right)}{(e_0 - \tilde{i}_\tau)!(n-e_0)!}$$

$$\times \left(\int_0^1 y^{\tilde{i}_\tau} \left(1-yF(\tau)\right)^{n-\tilde{i}_\tau} {}_2F_1\left(1, \tilde{i}_\tau - n, \tilde{i}_\tau + 2, \frac{-yF(\tau)}{1-yF(\tau)}\right) dy \right)^{-1} dp$$

Appendix D. Ordinary differential equations for the care home model

To model the background epidemic, we use the following system of ordinary differential equations

$$\frac{dS}{dt} = -\beta\frac{SI}{N},$$

$$\frac{dE}{dt} = \beta\frac{SI}{N} - \rho E,$$

$$\frac{dI}{dt} = \rho E - \gamma I,$$

$$\frac{dR}{dt} = \gamma I.$$

This provides a force of infection that is used to model seeding within each care home via the Sellke construction, with the details provided in the main text. Once infection has been seeded within a care home, the epidemic progresses using the following ordinary differential equations

$$\frac{dS}{dt} = -S\frac{(\beta_C I + r_P\beta_c P)}{N},$$

$$\frac{dE}{dt} = S\frac{(\beta_C I + r_P\beta_c P)}{N} - \rho E,$$

$$\frac{dP}{dt} = \rho E - \delta P,$$

$$\frac{dI}{dt} = \delta P - \gamma I,$$

$$\frac{dM}{dt} = \gamma p_M I,$$

$$\frac{dH}{dt} = \gamma p_H I - \eta H,$$

$$\frac{dR}{dt} = \eta p_R H,$$

$$\frac{dD}{dt} = \gamma p_D H,$$

$$\frac{dN}{dt} = -\gamma p_H I.$$

We add further compartments to the within care home model since we are interested in the different pathways that these individuals may take. For the background epidemic this is not important, since all we need is a force of infection provided by individuals in the infectious class.

Appendix E. Population and household transmission: Individual isolation and household quarantine

Individual isolation does not intimately involve the household and so we assume that a fraction α_I of symptomatic cases self-isolates and ceases transmission outside the household, meaning that we take the baseline but with

$$r_{s\to e}(t,\mathbf{Q}) = \Lambda(t) + \sum_{n=1}^{n_{\max}}\sum_{s=0}^{n}\sum_{e=0}^{(n-s)}\sum_{p=0}^{(n-s-e)}\sum_{i=0}^{(n-s-e-p)}\left(p\beta_p + (1-\alpha_I)i\beta_i\right)Q_{n,s,e,p,i}\ .$$

To capture the essential features of household quarantine, we need to add states to the dynamical variables. Let $Q_{n,s,e,p,i,\mathbf{f}}(t)$ be the proportion of households in the population at time t of size n, with s susceptibles, e exposed, p prodromal, and i symptomatic infectious individuals, and with vector of 'flags' $\mathbf{f}$ representing implementation of more complex interventions.

We now suppose that a fraction α_S of households start to isolate when there is at least one symptomatic case in the household, and stop isolation 14 days (on average) after the absence of symptoms in the household. This is modelled by having a flag $f = 0$ if the household is not isolating and $f = 1$ if it is. The dynamics become

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}Q_{n,s,e,p,i,1} = &-\left(er_{e\to p}+pr_{p\to i}+ir_{i\to\varnothing}+n^{-\eta}sp\tau_p+n^{-\eta}si\tau_i+1_{\{i=0\}}\sigma\right)Q_{n,s,e,p,i,1}\\ &+(e+1)r_{e\to p}Q_{n,s,e+1,p-1,i,1}+(p+1)r_{p\to i}Q_{n,s,e,p+1,i-1,1}\\ &+(i+1)r_{i\to\varnothing}Q_{n,s,e,p,i+1,1}+n^{-\eta}(s+1)p\tau_pQ_{n,s+1,e-1,p,i,1}\\ &+n^{-\eta}(s+1)i\tau_iQ_{n,s+1,e-1,p,i,1}+1_{\{i=1\&s+e+p=n-1\}}\alpha_S(p+1)r_{p\to i}Q_{n,s,e,p+1,i-1,0}\ ,\end{aligned}$$

for households in quarantine

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}Q_{n,s,e,p,i,0} = &-\left(sr_{s\to e}(t,\mathbf{Q})+er_{e\to p}+pr_{p\to i}+ir_{i\to\varnothing}+n^{-\eta}sp\tau_p+n^{-\eta}si\tau_i\right)Q_{n,s,e,p,i,0}\\ &+(s+1)r_{s\to e}(t,\mathbf{Q})Q_{n,s+1,e-1,p,i,0}+(e+1)r_{e\to p}Q_{n,s,e+1,p-1,i,0}\\ &+(p+1)r_{p\to i}Q_{n,s,e,p+1,i-1,0}+(i+1)r_{i\to\varphi}Q_{n,s,e,p,i+1,0}\\ &+n^{-\eta}(s+1)p\tau_pQ_{n,s+1,e-1,p,i,0}+n^{-\eta}(s+1)i\tau_iQ_{n,s+1,e-1,p,i,0}\\ &+1_{\{i=0\}}\sigma Q_{n,s,e,p,i,1}+\left(1-1_{\{i=1\&s+e+p=n-1\}}\alpha_S\right)(p+1)r_{p\to i}Q_{n,s,e,p+1,i-1,0}\ ,\end{aligned}$$

for households not in quarantine, with between-household term

$$r_{s\to e}(t,\mathbf{Q})=\Lambda(t)+\sum_{n=1}^{n_{\max}}\sum_{s=0}^{n}\sum_{e=0}^{(n-s)}\sum_{p=0}^{(n-s-e)}\sum_{i=0}^{(n-s-e-p)}\left(p\beta_p+i\beta_i\right)Q_{n,s,e,p,i,0}\ .$$

In this model, we assume that the duration of isolation after the absence of symptoms is exponentially distributed with mean equal to the fixed isolation period, given by $1/\sigma$. This assumption aids the modelling and is justifiable because, in reality, although a household may choose to isolate, they may not strictly follow the fixed period. It is likely that within-household isolation will increase transmission to other members of the household. We do not incorporate this property into the model, but this limitation is worth bearing in mind when drawing conclusions about the effectiveness of different strategies.

Author contributions

CO and HS compiled the manuscript. All authors were involved in the research and revising of the manuscript.

Data and materials

All data and code used in this analysis is provided in the Github repository https://github.com/thomasallanhouse/covid19-stochastics and https://github.com/thomasallanhouse/covid19-growth, with the exception of UK specific data. This data is provided by Public Health England under a data sharing agreement and we are unable to share this data.

References

Abramowitz, M., & Stegun, I. A. (1948). *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (Vol. 55). US Government printing office.

Alexander, H. K., & Bonhoeffer, S. (2012). Pre-existence and emergence of drug resistance in a generalized model of intra-host viral dynamics. *Epidemics, 4*(4), 187–202.

Andersson, H., & Britton, T. (2000). Stochastic epidemics in dynamic populations: Quasi-stationarity and extinction. *Journal of Mathematical Biology, 41*(6), 559–580.

Andreasen, V. (2011). The final size of an epidemic and its relation to the basic reproduction number. *Bulletin of Mathematical Biology, 73*(10), 2305–2321.

Ball, F. (1999). Stochastic and deterministic models for SIS epidemics among a population partitioned into households. *Mathematical Biosciences, 156*(1), 41–67.

Ball, F., Britton, T., House, T., Isham, V., Mollison, D., Pellis, L., et al. (2015). Seven challenges for metapopulation models of epidemics, including households models. *Epidemics, 10*, 63–67 (Challenges in Modelling Infectious Disease Dynamics).

Ball, F., Britton, T., & Sirl, D. (2011). Household epidemic models with varying infection response. *Journal of Mathematical Biology, 63*(2), 309–337.

Ball, F. G., Knock, E. S., & O'Neill, P. D. (2015). Stochastic epidemic models featuring contact tracing with delays. *Mathematical Biosciences, 266*, 23–35.

Ball, F., Mollison, D., & Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *Annals of Applied Probability*, 46–89.

Ball, F., Pellis, L., & Trapman, P. (2016). Reproduction numbers for epidemic models with households and other social structures II: Comparisons and implications for vaccination. *Mathematical Biosciences, 274*, 108–139.

Biggerstaff, M., Dahlgren, F. S., Fitzner, J., George, D., Hammond, A., Hall, I., et al. (2020). Coordinating the real-time use of global influenza activity data for better public health planning. *Influenza and Other Respiratory Viruses, 14*(2), 105–110.

Black, A. J., Geard, N., McCaw, J. M., McVernon, J., & Ross, J. V. (2017). Characterising pandemic severity and transmissibility from data collected during first few hundred studies. *Epidemics, 19*, 61–73.

Brauer, F. (2019). The final size of a serious epidemic. *Bulletin of Mathematical Biology, 81*(3), 869–877.

Britton, T., & Scalia Tomba, G. (2019). Estimation in emerging epidemics: Biases and remedies. *Journal of The Royal Society Interface, 16*(150), 20180670.

Butler, D. (2014). Models overestimate Ebola cases. *Nature, 515*(7525), 18.

Care Quality Commission. (2020). *CQC care directory - with ratings* (1 April 2020).

Cauchemez, S., Carrat, F., Viboud, C., Valleron, A. J., & Boëlle, P. Y. (2004). A bayesian MCMC approach to study transmission of influenza: Application to household longitudinal data. *Statistics in Medicine, 23*(22), 3469–3487.

Cauchemez, S., Donnelly, C. A., Reed, C., Ghani, A. C., Fraser, C., Kent, C. K., et al. (2009). *New England Journal of Medicine, 361*, 2619–2627.

C. Chew and G. Eysenbach. Pandemics in the age of twitter: Content analysis of tweets during the 2009 H1N1 outbreak. PloS One, 5(11), 2010.

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., et al. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science, 368*, 395–400.
Department of Health and Social Care. (2018). *SPI-M modelling summary for pandemic influenza, Edition*. Nov.
Fang, L., Karakiulakis, G., & Roth, M. (2020). Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? *The Lancet Respiratory Medicine, 8*(4), e21.
Farewell, V. T., Herzberg, A. M., James, K. W., Ho, L. M., & Leung, G. M. (2005). SARS incubation and quarantine times: When is an exposed individual known to be disease free? *Statistics in Medicine, 24*(22), 3431–3445.
Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., et al. (2020). *Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand*.
Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., et al. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science, 368*.
Funk, S., Gilad, E., Watkins, C., & Jansen, V. A. A. (2009). The spread of awareness and its impact on epidemic outbreaks. *Proceedings of the National Academy of Sciences, 106*(16), 6872–6877.
Ganyani, P., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J., et al. (2020). *Estimating the generation interval for COVID-19 based on symptom onset data*. medRxiv.
Goldstein, E., Paur, K., Fraser, C., Kenah, E., Wallinga, J., & Lipsitch, M. (2009). Reproductive numbers, epidemic spread and control in a community of households. *Mathematical Biosciences, 221*(1), 11–25.
Gostic, K. M., Gomez, A. C. R., Mummah, R. O., Kucharski, A. J., & Lloyd-Smith, J. O. (2020). Estimated effectiveness of symptom and risk screening to prevent the spread of COVID-19. *eLife, 9*, 1–18.
Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine, 382*, 1708–1720.
Hale, T., et al. (2020). Oxford COVID-19 government response tracker (OxCGRT). https://www.bsg.ox.ac.uk/research/research-projects/oxford-covid-19-government-response-tracker.
House, T., & Keeling, M. J. (2008). Deterministic epidemic models with explicit household structure. *Mathematical Biosciences, 213*(1), 29–39.
House, T., & Keeling, M. J. (2011). Epidemic prediction and control in clustered populations. *Journal of Theoretical Biology, 272*(1), 1–7.
Jack, S. (2020). *Coronavirus: GSK and Sanofi join forces to create vaccine*.
Kalbfleisch, J. D., & Lawless, J. F. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica*, 19–32.
Keeling, M. J., & Rohani, P. (2011). *Modeling infectious diseases in humans and animals*. Princeton University Press.
Keju, W. (2019). *China braces for world's biggest travel rush around Spring Festival*.
Kendall, D. G., et al. (1948). On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics, 19*(1), 1–15.
Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London - Series A: Containing Papers of a Mathematical and Physical Character, 115*(772), 700–721.
Kinyanjui, T., Middleton, J., Güttel, S., Cassell, J., Ross, J., & House, T. (2018). Scabies in residential care homes: Modelling, inference and interventions for well-connected population sub-units. *PLoS Computational Biology, 14*(3), Article e1006046.
Kraemer, M. U. G., Yang, C., Gutierrez, B., Wu, C., Klein, B., Pigott, D. M., et al. (2020). The effect of human mobility and control measures on the covid-19 epidemic in China. *Science, 368*(6490), 493–497.
Lau, M. S. Y., Dalziel, B. D., Funk, S., McClelland, A., Tiffany, A., Riley, S., et al. (2017). Spatial and temporal dynamics of superspreading events in the 2014-2015 West Africa Ebola epidemic. *Proceedings of the National Academy of Sciences of the United States of America, 114*(9), 2337–2342.
Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H., et al. (2020). *The incubation period of 2019-nCoV from publicly reported confirmed cases: Estimation and application*. medRxiv.
Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S., et al. (2020). incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine, 9*(2), 538.
Lesko, J. K., Keil, A. P., & Edwards, J. K. (2020). The epidemiologic toolbox: Identifying, honing, and using the right tools for the job. *American Journal of Epidemiology*, kwaa030.
Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., et al. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science, 368*, 489–493.
Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine, 27*.
Magpantay, F. M. G., Riolo, M. A., Domenech de Cellès, M., King, A. A., & Rohani, P. (2014). Epidemiological consequences of imperfect. *SIAM Journal on Applied Mathematics, 74*(6), 1810–1830.
Mahase, E. (2020). China coronavirus: What do we know so far? *BMJ, 368*.
Majumder, M. S., & Mandl, K. D. (2020). Early in the epidemic: Impact of preprints on global discourse about COVID-19 transmissibility. *The Lancet Global Health, 8*.
Mizumoto, K., Kagaya, K., Zarebski, A., & Chowell, G. (2020). Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan. *Euro Surveillance, 25*(10), 2000180, 2020.
Morris, H.. The largest human migration on the planet: What happens in China when 1.4bn go on holiday at the same time. https://www.telegraph.co.uk/travel/news/chinese-new-year-chunyun-in-numbers/.
Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., et al. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine, 5*(3).
Nishiura, H. (2010). Time variations in the generation time of an infectious disease: Implications for sampling to appropriately quantify transmission potential. *Mathematical Biosciences and Engineering, 7*(4), 851–869.
Nishiura, H., Kobayashi, T., Miyama, T., Suzuki, A., Jung, S., Hayashi, K., et al. (2020). *Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19)*. medRxiv.
Office for National Statistics. (2011). *2001 census aggregate data, Edition*. May.
Office for National Statistics. (2019). *Families and households, edition*, 15 November https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/datasets/familiesandhouseholdsfamiliesandhouseholds.
Pellis, L., Ball, F., & Trapman, P. (2012). Reproduction numbers for epidemic models with households and other social structures. I. definition and calculation of R_0. *Mathematical Biosciences, 235*(1), 85–97.
Pellis, L., Cauchemez, S., Ferguson, N. M., & Fraser, C. (2020). Systematic selection between age and household structure for models aimed at emerging epidemic predictions. *Nature Communications, 11*(1).
Pellis, L., Ferguson, N. M., & Fraser, C. (2011). Epidemic growth rate and household reproduction number in communities of households, schools and workplaces. *Journal of Mathematical Biology, 63*(4), 691–734.
Pellis, L., Scarabel, F., Stage, H. B., Overton, C. E., Chappell, L. H. K., Lythgoe, K. A., et al. (2020). *Challenges in control of covid-19: Short doubling time and long delay to effect of interventions*. arXiv.
Recorded daily case updates of covid-19 cases. Data scraped daily from 21-1-2020 to 3-2-2020 http://3g.dxy.cn/newh5/view/pneumonia.
Remuzzi, A., & Remuzzi, G. (2020). COVID-19 and Italy: What next? *The Lancet, 395*.
Rimmer, A. (2020). COVID-19: Disproportionate impact on ethnic minority healthcare workers will be explored by government. *BMJ, 369*.
Ross, R. (1910). *The prevention of malaria*. Dutton.

Ross, J. V., House, T., & Keeling, M. J. (2010). Calculation of disease dynamics in a population of households. *PloS One, 5*(3), Article e9666.
Rubin, G. J., Amlôt, R., Page, L., & Wessely, S. (2009). Public perceptions, anxiety, and behaviour change in relation to the swine flu outbreak: Cross sectional telephone survey. *BMJ, 339*, b2651.
Scalia Tomba, G., Svensson, Å., Asikainen, T., & Giesecke, J. (2010). Some model based considerations on observing generation times for communicable diseases. *Mathematical Biosciences, 223*(1), 24–31.
Sellke, T. (1983). On the asymptotic distribution of the size of a stochastic epidemic. *Journal of Applied Probability, 20*(2), 390–394.
Shaw, L. M. (2016). *SIR epidemics in a population of households*. PhD thesis. University of Nottingham.
Sun, J. (1995). Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies. *Biometrics*, 1096–1104.
Su, Y., & Wang, J. (2012). Modeling left-truncated and right-censored survival data with longitudinal covariates. *Annals of Statistics, 40*(3), 1465.
Sun, K., Chen, J., & Viboud, C. (2020). Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: A population-level observational study. *The Lancet Digital Health, 2*.
Svensson, Å. (2007). A note on generation times in epidemic models. *Mathematical Biosciences, 208*(1), 300–311.
Taylor, D. J., Weaver, M. A., & Roddy, R. E. (2003). Evaluating factors associated with STD infection in a study with interval-censored event times and an unknown proportion of participants not at risk for disease. *Statistics in Medicine, 22*(13), 2191–2204.
Valdano, E., Poletto, C., Boelle, P., & Colizza, V. (2019). *Reorganization of nurse scheduling reduces the risk of healthcare associated infections.* medRxiv.
Van Kerkhove, M. D., & Ferguson, N. M. (2012). Epidemic and intervention modelling: A scientific rationale for policy decisions? Lessons from the 2009 influenza pandemic. *Bulletin of the World Health Organization, 90*, 306–310.
Varia, M., Wilson, S., Sarwal, S., McGeer, A., Gournis, E., Galanis, E., et al. (2003). Investigation of a nosocomial outbreak of severe acute respiratory syndrome (SARS) in Toronto, Canada. *Canadian Medical Association Journal, 169*(4), 285–292.
Virlogeux, V., Fang, V. J., Park, M., Wu, J. T., & Cowling, B. J. (2016). Comparison of incubation period distribution of human infections with MERS-CoV in South Korea and Saudi Arabia. *Scientific Reports, 6*(1), 35839.
Wilder-Smith, A., Chiew, C. J., & Lee, V. J. (2020). Can we contain the COVID-19 outbreak with the same measures as for SARS? *The Lancet Infectious Diseases, 20*.
World Health Organisation. (2020). *Coronavirus disease 2019 (COVID-19). Situation report - 69*.
Wu, J. T., Leung, K., Bushman, M., Kishore, N., Niehus, R., de Salazar, P. M., et al. (2020). Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nature Medicine*, 1–5.
Wu, Z., & McGoogan, J. M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72 314 cases from the Chinese center for disease control and prevention. *Jama, 323*, 1239–1242.
Yang, J., Zheng, Y., Gou, X., Pu, K., Chen, Z., Guo, Q., et al. (2020). Prevalence of comorbidities in the novel Wuhan coronavirus (COVID-19) infection: A systematic review and meta-analysis. *International Journal of Infectious Diseases, 94*.
Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *The Lancet, 395*.