

PUBLISHED VERSION

Andrew J. Black, Nicholas Gear, James M. McCaw, Jodie McVernon, Joshua V. Ross
Characterising pandemic severity and transmissibility from data collected during first few
hundred studies

Epidemics, 2017; 19:61-73

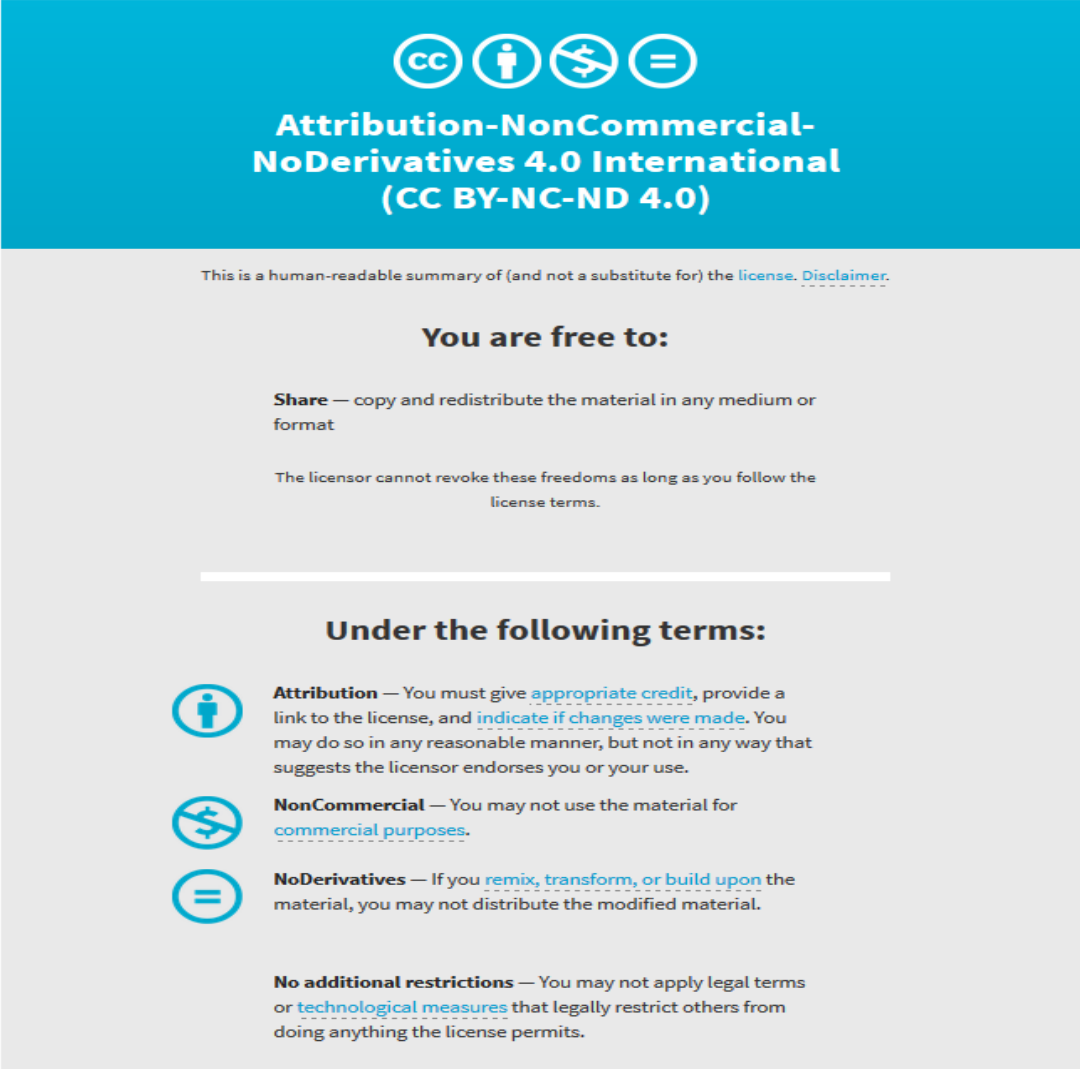
© 2017 The Author(s). Published by Elsevier B.V. This is an open access article under the CCBY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Originally published at:

<http://doi.org/10.1016/j.epidem.2017.01.004>

PERMISSIONS

<http://creativecommons.org/licenses/by-nc-nd/4.0/>



The image shows the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license graphic. It features a blue header with the license name and icons for Attribution (person), Non-Commercial (dollar sign with slash), and No Derivatives (equals sign). Below the header, it states: "This is a human-readable summary of (and not a substitute for) the [license](#). [Disclaimer](#)." The main content is on a light gray background and is divided into two sections: "You are free to:" and "Under the following terms:". Under "You are free to:", it lists "Share" — copy and redistribute the material in any medium or format, and notes that the licensor cannot revoke these freedoms as long as you follow the license terms. Under "Under the following terms:", it lists three conditions: "Attribution" — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. "NonCommercial" — You may not use the material for [commercial purposes](#). "NoDerivatives" — If you [remix, transform, or build upon](#) the material, you may not distribute the modified material. At the bottom, it states "No additional restrictions" — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

21 September 2017

<http://hdl.handle.net/2440/107344>



Characterising pandemic severity and transmissibility from data collected during first few hundred studies



Andrew J. Black^{a,g,*}, Nicholas Geard^{b,c}, James M. McCaw^{b,d,e}, Jodie McVernon^{b,e,f}, Joshua V. Ross^{a,g}

^a School of Mathematical Sciences, The University of Adelaide, Adelaide, SA 5005, Australia

^b Center for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC 3010, Australia

^c School of Computing and Information Systems, Melbourne School of Engineering, The University of Melbourne, Melbourne, VIC 3010, Australia

^d School of Mathematics and Statistics, The University of Melbourne, Melbourne, VIC 3010, Australia

^e Murdoch Childrens Research Institute, Royal Childrens Hospital, VIC, Australia

^f The Peter Doherty Institute for Infection and Immunity, The University of Melbourne and Royal Melbourne Hospital, Melbourne, VIC 3000, Australia

^g ACEMS, School of Mathematical Sciences, University of Adelaide, Adelaide, SA 5005, Australia

ARTICLE INFO

Article history:

Received 5 May 2016

Received in revised form 9 January 2017

Accepted 15 January 2017

Available online 19 January 2017

Keywords:

Pandemic

Influenza

Households

Markov chain

Parameter inference

ABSTRACT

Early estimation of the probable impact of a pandemic influenza outbreak can assist public health authorities to ensure that response measures are proportionate to the scale of the threat. Recently, frameworks based on transmissibility and severity have been proposed for initial characterization of pandemic impact. Data requirements to inform this assessment may be provided by “First Few Hundred” (FF100) studies, which involve surveillance—possibly in person, or via telephone—of household members of confirmed cases. This process of enhanced case finding enables detection of cases across the full spectrum of clinical severity, including the date of symptom onset. Such surveillance is continued until data for a few hundred cases, or satisfactory characterization of the pandemic strain, has been achieved.

We present a method for analysing these data, at the household level, to provide a posterior distribution for the parameters of a model that can be interpreted in terms of severity and transmissibility of a pandemic strain. We account for imperfect case detection, where individuals are only observed with some probability that can increase after a first case is detected. Furthermore, we test this methodology using simulated data generated by an independent model, developed for a different purpose and incorporating more complex disease and social dynamics. Our method recovers transmissibility and severity parameters to a high degree of accuracy and provides a computationally efficient approach to estimating the impact of an outbreak in its early stages.

© 2017 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Influenza pandemics occur following the emergence of a new strain of the influenza virus; a strain that is sufficiently immunologically distinct to previous strains such that the majority of the population has negligible levels of immunity against it. Past influenza pandemics have given rise to dramatically different scales of impact; the 1918 Spanish influenza pandemic has been estimated to have caused approximately 40 million deaths worldwide, whereas the 2009 Swine Flu pandemic has been estimated to have caused approximately 14,000 deaths worldwide. The ability to assess the expected impact as early as possible following the

emergence of a new strain is of obvious benefit to informing proportionate public health response efforts (Van Kerkhove et al., 2010; Van Kerkhove and Ferguson, 2012; McCaw et al., 2013).

The benefits of early assessment, and the dependency of response plans and actions hinging on the characterisation of the pandemic strain, has led to the development of response frameworks based on the transmissibility and severity of a pandemic (McCaw et al., 2013; Reed et al., 2013; Australian Department of Health, 2014; Riley et al., 2015). The motivation is based upon these two factors—severity and transmissibility—being strong determinants of impact: severity moderates impact through illness, demand on health services and potential deaths, and transmissibility influences the speed of spread, timing of peak demand on health services and the overall extent of the pandemic. Transmissibility also determines the likely impact of interventions; often it is possible to estimate the proportion of transmission that an

* Corresponding author.

E-mail address: andrew.black@adelaide.edu.au (A.J. Black).

intervention might avert, hence allowing the estimation of the possibility of containment or of the reduction in attack rate. A number of studies will be required in the initial stages of a pandemic to make a rigorous characterisation of the emergent strain. Enhanced case finding efforts directed at contacts of early identified cases, also known as “First Few Hundred” (FF100) studies, provide rich information on disease characterisation and spread (Health Protection Agency England, 2009; Ghani et al., 2009; Cauchemez et al., 2009; McLean et al., 2010; van Gageldonk-Lafeber et al., 2012; Australian Department of Health, 2014).

An FF100 study, as the name suggests, involves recording data on the first few hundred cases, early in the pandemic. The most well known design is from the UK (Health Protection Agency England, 2009): following the first confirmed case of the pandemic strain, that individual and all other members of their household are surveilled—possibly in person, or via telephone—to identify day(s) of symptom onset and disease characteristics in other household members. Supplementary information concerning the household, such as household size, and possibly age composition, are also recorded. Studies are continued until data for a few hundred cases, enabling satisfactory characterisation of the pandemic strain, has been collected. For this study we assume that household sizes and dates of symptom onset of members of households, up to the first few hundred cases, are available. The base scenario we consider is one of *partial* detection, where each infectious individual is only observed with some probability.

In this paper we develop a novel methodology for analysing and performing inference on this partially observed, FF100 type, household level data. The assumed underlying model of transmission dynamics is a Markovian households model where there exists two-levels of mixing—within-households and between-households (Ball et al., 1997; Black et al., 2013). When analysing data, we make the assumption that there is only a single introduction of infection into a household. Essentially this means we perform inference on a large number of small independent outbreaks rather than a single larger outbreak (O’Dea et al., 2014). Our detection model accounts for asymptomatic cases as well as imperfect surveillance. Cases are initially detected with some probability that can then increase after the first detection. This increase of the detection probability is due to the increased surveillance of a household after the first case detection as appropriate for an FF100 study. Previous studies have used household data for inference (Cauchemez et al., 2004, 2009; Ghani et al., 2009; Lau et al., 2015), but generally only for estimating secondary attack rates. To analyse time series data and allow for estimates of transmission rates requires a completely mechanistic model as we adopt herein. Additionally the two main determinants of impact in the early stages of a pandemic, transmissibility and severity (McCaw et al., 2013; Reed et al., 2013), are simply determined from our model.

For inference, we implement a Bayesian Markov chain Monte Carlo (MCMC) scheme with exact evaluation of the likelihood for all the observed data. Exact likelihood evaluation is made possible through optimisation of code based upon probabilistic arguments and a novel data structure for minimising the computations required. This approach provides a posterior distribution over the parameters of the model that can then be interpreted in terms of the severity and transmissibility of a pandemic strain. The only other method for inference with such data is that of multiple imputation or data augmentation (Gibson and Renshaw, 1998; O’Neill and Roberts, 1999; Cauchemez et al., 2004; Lau et al., 2015). In this approach, all unobserved events are treated as unknowns to also be inferred within the MCMC routine, which allows a great deal of flexibility in modelling. The trade off of such an approach is that the MCMC scheme needed to sample from the joint distribution of parameters and unknown data is more complex and convergence can be an issue when there is a large amount of missing data to be

inferred (McKinley et al., 2014). Such an approach is quite different to that adopted in this paper where we essentially consider all paths of the process at once for a given set of parameters, allowing us to efficiently scale the algorithm.

The efficiency of the method is important as it allows us to perform inference on many, and very large, data sets. This in turn allows a proper quantification of the variability inherent to this sort of study, to a degree not previously achieved. In any outbreak there is a large amount of inherent randomness, but this is magnified in FF100 studies due to the small size of typical households and partial observation. We demonstrate correct convergence of the estimates as the amount of data is increased, but more importantly study what bias is introduced by smaller, realistic size, data. Finally the efficiency of our method also ensures utility in real-time during an enacted FF100 study, including timely advice as to *when* enhanced surveillance (i.e., FF100 studies) can be stopped due to sufficient acquisition of data. Furthermore, our methodology provides a way forward to investigate variations on the FF100 study design and their effectiveness for determining transmissibility and severity for a range of potential pandemic scenarios.

A difficulty with methodology for pandemics, and in particular FF100 studies, is a lack of datasets both due to infrequent pandemic occurrence and the relatively new consideration of FF100 studies. This makes validation of any proposed methodology difficult. Whilst one may, as we undertake herein, test the methodology on data simulated from the underlying model upon which the methodology is developed, this does not typically provide adequate assurance that the methodology will be sufficiently accurate in the event of the next pandemic, where almost certainly the modelling assumptions will be violated. Here we make an attempt to provide some assurance. This is achieved by testing our methodology using data that is produced by an independent model, a model that has been developed for a different purpose and that should more accurately reflect true pandemic (and social) dynamics. The particular model we use herein is the microsimulation model of Geard et al. (2013, 2015), calibrated for a pandemic influenza scenario.

2. Methods

We first describe the stochastic households model that incorporates partial detection of cases that we use to perform inference. Next we describe the form of the data we assume and how we structure it before detailing how to calculate the likelihood for observations from a single household. This allows us to develop the theory without the complications of making it efficient for inference over multiple households; this is done in the next section. We describe how the data from FF100 studies can be naturally described using a tree structure and then we give an algorithm to calculate the likelihood using this approach. Finally we discuss how data are generated for validation of the methodology. The first validation is performed using data simulated from the same model assumed for inference. The second validation is performed on data generated by a different, more complex model. Brief details of this are given, emphasising the differences between the two models.

2.1. Stochastic household model

The epidemic dynamics within a household are modelled with a continuous-time Markov chain. To facilitate efficient inference, we make the assumption that there is only one introduction into any household that experiences infection. This is likely to be plausible in the early stages of a pandemic. Note that this is not an assumption in the micro-simulation model used to generate data for validation of our methods. Thus we can assess this assumption and its implications for inference.

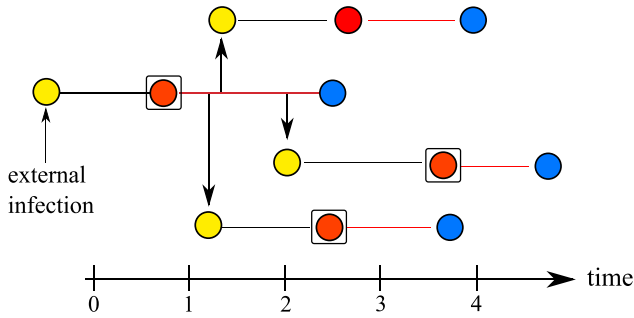


Fig. 1. A typical household epidemic. Individuals (represented as circles) become exposed (yellow) following infection transmission, become infectious (red) which is when they can also be detected, and finally recover (blue). Only the events surrounded by the black boxes are observed, the rest are not observed. Thus in this example, there is one infectious individual who is undetected on day 2, while the remaining three are detected. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The base model we use for inference is an SE_2I_2R model where both the latent and infectious periods are split into two stages, giving them an Erlang distribution with mean periods $1/\sigma$ and $1/\gamma$ respectively. The number of stages can be changed (Black and Ross, 2013) but this model has been used for previous influenza studies (Baguelin et al., 2010; House et al., 2011; Black et al., 2013) and represents a good trade-off between realism—influenza had a high variance in the next generation distribution that is accurately captured by Erlang distributions (Donnelly et al., 2010)—and computational efficiency.

The classes an individual can be in are: S, E_1, E_2, I_1, I_2 and R , and the size of the household, N , is fixed. Transmission within the household is assumed to be frequency dependent with rate β , thus β is independent of the size of the household and the overall rate of infection is $\beta S(I_1 + I_2)/(N - 1)$ (McCallum et al., 2001). As in the microsimulation described in Section 2.6, we assume that symptoms coincide with the onset of infectiousness and that there is a per-case probability of detection of an infected individual. This is incorporated into the model by splitting the transition $E_2 \rightarrow I_1$ into two parts: with probability ρ the event is detected and, with probability $(1 - \rho)$ it goes undetected (a typical scenario is illustrated in Fig. 1). We assume that the probability of initial detection in the household is $\rho = p$, and afterwards, once FF100 surveillance has been enacted, $\rho = q$ where $q \geq p$.

The two main parameters we wish to estimate for a given outbreak are the transmissibility and severity (McCaw et al., 2013; Reed et al., 2013; Riley et al., 2015). Transmissibility is quantified by the expected household secondary attack rate (hSAR), which is the proportion of individuals infected by an index case in a household of a given size. In our model the hSAR can be calculated exactly from the within-household epidemic final size distribution, which depends only on the ratio β/γ and the household size, N (Ball, 1986). As it depends on N , when we summarise our results later we plot posteriors in terms of the expected hSAR for a household of size $N=3$, denoted $hSAR_3$. Similar plots could be made for any household size, but the posteriors scale in the same way with increased data. We quantify severity by the detection probability q , that is the probability of a case detection with enhanced surveillance in place. This reflects that more significant presentation of symptoms, or even hospitalisations, will result in a higher probability of detection of cases.

Instead of keeping track of the population numbers within the household—i.e., the numbers of S, E_1 etc.—we instead specify the process in terms of the numbers of events of a given type that have occurred. This is known as the degree-of-advancement (DA) or reaction-count representation of the stochastic process (van Kampen, 1992; Sunkara, 2009; Jenkinson and Goutsias, 2012;

Black and Ross, 2015). As the counts only ever increase, this representation simplifies some aspects of calculating the likelihood. Adopting this representation of the stochastic process along with a lexicographical ordering of the state space also allows us to calculate the probability mass function of the state of the chain using computationally-efficient methods (Jenkinson and Goutsias, 2012; Black and Ross, 2015).

We introduce the variables $Z_i(t), i=1, \dots, 6$, which count the events of each type which have occurred prior to time t . These events, their transitions and rates are summarised in Table 1. The state of the system is $\mathbf{Z}(t)=(Z_1(t), \dots, Z_6(t))$. The relation between $Z_i(t)$ and the population variables is (dropping the dependence on t),

$$\begin{aligned} S &= N - Z_1, \\ E_1 &= Z_1 - Z_2, \\ E_2 &= Z_2 - Z_3 - Z_4, \\ I_1 &= Z_3 + Z_4 - Z_5, \\ I_2 &= Z_5 - Z_6, \\ R &= Z_6, \end{aligned} \tag{1}$$

and the cumulative number of detected cases within the household is counted by Z_3 . Note that the variable Z_1 also counts the initial infection into the household and hence sets the initial condition for the system, i.e., $\mathbf{Z}(0)=(1, 0, 0, 0, 0, 0)$. The relations in (1) allow us to write the rates of each transition in terms of Z_i and are given in Table 1.

The state space of the process is

$$\mathcal{S} = \left\{ (Z_1, \dots, Z_6) \in (\mathbb{Z}^+)^6 : 0 \leq Z_6 \leq Z_5 \leq Z_3 + Z_4 \leq Z_2 \leq Z_1 \leq N \right\}. \tag{2}$$

This is partitioned into two subsets, $\mathcal{S} = A \cup C$, where A are absorbing states and C are transient states. Absorbing states are those where the outbreak has ended, i.e., the number of infection events equals the number of recovery events ($Z_1 = Z_6$). As we do not consider more than one introduction into a household, an epidemic always leads to absorption of the Markov chain. The mapping between states of the system and event counts is useful in specifying various subsets of \mathcal{S} . We do this with the matrix

$$\mathcal{Z} = (z_{ij}) \quad i \in \mathcal{S}, \quad j = 1, \dots, 6, \tag{3}$$

where z_{ij} is the number of events of type j associated with state i . The rows of \mathcal{Z} , which we define as $\mathbf{z}_i = (z_{i1}, \dots, z_{i6})$, can then be used to index states of the system. For example, the absorbing states, A , can be specified as

$$A = \{i \in \mathcal{S} | z_{i1} = z_{i6}\}. \tag{4}$$

We order the state space by detections (Z_3) first so the generator (transition rate matrix) Q is partitioned into blocks indexed by Z_3 . The state space is enumerated, and hence the elements of the matrix \mathcal{Z} (see (3)) are determined, by using a system of nested loops which iterate over the variables Z_1 to Z_6 . The ordering of \mathcal{S} means that the generator, Q , is triangular which is required for the efficiency of the algorithms described later. The generator is sparse and can be written as a sum of matrices multiplied by the various parameters of the model,

$$\begin{aligned} Q &= \beta Q_1 + 2\sigma Q_2 + 2p\sigma Q_3 + 2(1-p)\sigma Q_4 + 2q\sigma Q_5 \\ &\quad + 2(1-q)\sigma Q_6 + 2\gamma Q_7 + 2\gamma Q_8. \end{aligned} \tag{5}$$

The matrices $Q_i, i=1, \dots, 8$ can be pre-calculated for a given household size, N , as their structure does not change. Thus forming the

Table 1
Transitions, event counters and rates defining the SE_2I_2R partial detection model. Initially, $\rho=p$, then after the first infection $\rho=q$, where $q \geq p$, representing an increased chance of detection due to surveillance of the household.

Event	Transition	Counter (+1)	Rate
Infection	$S \rightarrow E_1$	Z_1	$\beta(N - Z_1)(Z_3 + Z_4 - Z_6)/(N - 1)$
Latent progression	$E_1 \rightarrow E_2$	Z_2	$2\sigma(Z_1 - Z_2)$
Become infectious (detected)	$E_2 \rightarrow I_1$	Z_3	$2\rho\sigma(Z_2 - Z_3 - Z_4)$
Become infectious (undetected)	$E_2 \rightarrow I_1$	Z_4	$2(1 - \rho)\sigma(Z_2 - Z_3 - Z_4)$
Infectious progression	$I_1 \rightarrow I_2$	Z_5	$2\gamma(Z_3 + Z_4 - Z_5)$
Recovery	$I_2 \rightarrow R$	Z_6	$2\gamma(Z_5 - Z_6)$

matrix Q for any given set of parameters can be done efficiently. The dynamics of the system are given by the forward (master) equation,

$$\frac{d\phi(t)}{dt} = \phi(t)Q, \tag{6}$$

where $\phi_i(t) = \Pr(\mathbf{Z}(t) = \mathbf{z}_i)$ is the probability that the system is in state i at time t .

2.2. Data

We assume that the data available from an FF100 study is of the form of time-series of detection counts stratified by household and binned into days, where we define day t as the interval $(t - 1, t]$ for $t \geq 1$. Two realisations, generated by the microsimulation described in Section 2.6, with (a) high severity ($p=0.5, q=0.9$) and (b) low severity ($p=0.1, q=0.5$) are shown in Fig. 2. These outputs show the times of new detections (red dots) and the times of all other cases that have gone undetected (grey dots). The area of the dots represents the number of new detected cases on a given day. As these are stochastic, and households are small, there is a reasonable amount of variability present, particularly in the early stages of the outbreak (Black et al., 2014). We can clearly see that in the low severity scenario, many of the early cases are missed, so for a given number of households we have less data for inference.

We can perform inference on these data in two ways, either assuming we have observed completed outbreaks for a given number of households (essentially assuming that all households have been observed for a long enough period such that the outbreak is over), or that we have observed the beginning of an outbreak up to some time horizon. In the second scenario we will have a mixture of data, some from households where the outbreak is over and others where it is potentially ongoing. At the end of each day a new posterior can be calculated incorporating new data from that day. Note that when an outbreak is still ongoing, null days, where no further cases are detected in a household, still contribute information. Note that it is also possible to analyse the complete case data (detected plus undetected cases) using our model by setting $p=q=1$. This allows us to perform inference for the underlying epidemiological parameters, independent of the observational process, which will be useful when analysing data from the microsimulation model later.

As the outbreaks are assumed to be independent the time for each can be scaled to start at zero; thus, for a given household, the first detected case(s) always fall within day 1. The time series for the j 'th household is then represented by a vector, $\mathbf{d}^{(j)} = (d_i)^j$, where d_i is the cumulative number of detections, calculated at discrete time points $t_i = 1, \dots, n$, which correspond to the end of each day after the first detection is made. For example in Fig. 1, $\mathbf{d} = (1, 1, 2, 3)$ represents a single detection on day one, then further detections on days three and four. Surveillance of a household can be carried out indefinitely, giving a time series of any length, but in practice we can truncate \mathbf{d} after a particular number of days where the state has not changed, either because there are no more infectious individuals or

we have failed to detect later events. This truncation then allows for more efficient inference. For example,

$$(1, 1, 2, 2, 2, 2, 2, 2) \rightarrow (1, 1, 2).$$

If a time series is truncated, then this must be recorded for the purposes of calculating the likelihood. Thus we define the variable b_j for the j 'th household such that

$$b_j = \begin{cases} 1 & \text{if } \mathbf{d}^{(j)} \text{ is truncated} \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

A non-truncated time series represents a potentially ongoing outbreak.

2.3. Likelihood for a single household

In this section we develop the theory needed to calculate the likelihood of observing a given time series of detection events within a single household, \mathbf{d} (dropping the index j for now). This allows us to give a relatively simple exposition of the theory and our methodology before we describe how this calculation can be made more efficient when we wish to perform inference on a large number of time series simultaneously.

Let $\theta = (\beta, \sigma, \gamma, p, q)$ be the vector of parameters we wish to estimate. The likelihood of observing $\mathbf{d} = (d_i)_{1:n}$ in a household of a given size is then

$$L(\mathbf{d}; \theta) = P(Z_3(t_1) = d_1) \prod_{i=2}^n P(Z_3(t_i) = d_i | Z_3(t_{i-1}) = d_{i-1}). \tag{8}$$

The exact conditional probabilities in the likelihood can be evaluated iteratively from the dynamics of the Markov chain, which are given by Eq. (6). Let $\mathbb{1}_{S'} : S \rightarrow \{0, 1\}$ be an indicator vector of the subset S' of the state space S such that

$$\mathbb{1}_{S'}(x) = \begin{cases} 1 & \text{if } x \in S' \\ 0 & \text{if } x \notin S' \end{cases}. \tag{9}$$

The conditional probability terms in Eq. (8) can be written as,

$$P(Z_3(t_i) = d_i | Z_3(t_{i-1}) = d_{i-1}) = \mathbb{1}_{\{i \in S | Z_{i3} = d_i\}} \cdot \phi(t_i | \mathbf{d}_{1:i-1}) \tag{10}$$

which is a summation of the elements of $\phi(t_i | \mathbf{d}_{1:i-1})$ where $Z_3 = d_i$. The distribution, conditioned on the observations up to and including d_i will be

$$\phi(t_i | \mathbf{d}_{1:i}) = \frac{\phi(t_i | \mathbf{d}_{1:i-1}) \odot \mathbb{1}_{\{i \in S | Z_{i3} = d_i\}}}{\phi(t_i | \mathbf{d}_{1:i-1}) \cdot \mathbb{1}_{\{i \in S | Z_{i3} = d_i\}}}, \tag{11}$$

where \odot denotes an element-wise vector product. The operation of Eq. (11) essentially sets the elements of $\phi(t_i | \mathbf{d}_{1:i-1})$ corresponding to states with $Z_3 \neq d_i$ equal to zero, with the denominator normalising that result so that the resulting distribution sums to 1. Note that the procedure above is a modified version of the forward filtering step of the forward-backward algorithm for hidden Markov models (Sarkka, 2013).

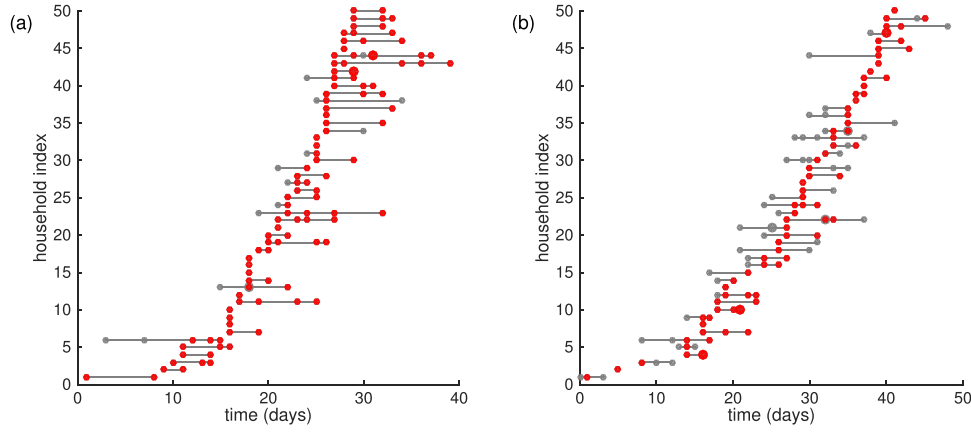


Fig. 2. Case times for the first 50 households. Data generated from the microsimulation model (described later) showing high severity (a) and low severity (b) outbreaks. Households are ordered by the time of the first detection. Red dots mark the days on which new cases are detected and grey dots mark times of undetected cases. The area of the dots indicates the number on that day. Only households with at least 1 detected case are shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The distribution $\phi(t_i|\mathbf{d}_{1:i-1})$ is calculated from forward integration of the distribution at the end of the previous day, $\phi(t_{i-1}|\mathbf{d}_{1:i-1})$, using Eq. (6). In this paper we evaluate (6) using an implicit Euler method (Jenkinson and Goutsias, 2012). The probability mass function, ϕ , then obeys systems of linear equations,

$$\phi(s + \tau)(1 - \tau Q) = \phi(s) \quad (12)$$

where τ is the time step. Due to the triangular structure of Q , Eq. (12) can be solved via forward substitution. The accuracy of the solution depends only on the size of the time-step, τ (Jenkinson and Goutsias, 2012), which is taken as 10^{-2} in this paper. Thus to integrate the dynamics over an observational time step of one day, Eq. (12) must be recursively solved $1/\tau = 100$ times.

Given an initial condition, the above procedure of integrating forward and conditioning can be carried out iteratively. As we scale the problem so that the first detection(s) occur in the first day, $t = (0, 1]$, the initial condition that we need is $\phi(1|d_1)$, the distribution of the process at the end of the first day, conditioned on having observed d_1 cases in that day. In calculating this we need to account for the fact that the first detection may not be the first actual case and that we do not know the exact time of the first case, only that it happened at some time within the day. To calculate the distribution at $t = 1$, we assume that there is a uniform probability of the initial detection event occurring within the interval $(0, 1]$ and thus the state of the system at the end of the day is given by the integral,

$$\phi(1) = \int_0^1 \phi^* e^{Q(1-s)} ds, \quad (13)$$

where ϕ^* is the distribution of the process at the time of the first detection. We first discuss how we evaluate this integral and then how to calculate ϕ^* . Note that Eq. (13) is the solution of the differential equations

$$\frac{d\phi(t)}{dt} = \phi(t)Q + \phi^* \quad (14)$$

evaluated at $t = 1$ with initial condition $\phi(0) = \mathbf{0}$. Integrating Eq. (14) from u to $u + \tau$ yields,

$$\phi(u + \tau) - \phi(u) = \int_u^{u+\tau} \phi(s)Q + \phi^* ds.$$

Approximating the integral using a rectangle method with height such that the top right corner matches the function, we obtain

$$\phi(u + \tau) - \phi(u) = \tau Q \phi(u + \tau) + \tau \phi^*.$$

Rearranging this we have a system of linear equations that can be solved by forward substitution,

$$(1 - \tau Q)\phi(u + \tau) = \phi(u) + \tau \phi^*. \quad (15)$$

The distribution at the end of day one, $\phi(1)$, is then calculated by recursively solving (15) starting from $u = 0$ with $\phi(0) = \mathbf{0}$. Once $\phi(1)$ is calculated, the first term in the likelihood is calculated as,

$$P(Z_3(t_1) = d_1) = \mathbb{1}_{\{i \in S | z_{i3} = d_1\}} \cdot \phi(1) \quad (16)$$

and the vector is conditioned on the observation using Eq. (11) to yield $\phi(1|d_1)$ —the initial condition for the iterative procedure detailed above.

We can calculate the distribution of the process at the time of the first detection, ϕ^* , from the corresponding hitting probabilities, i.e. the probability of hitting a state i , given that the system starts in state j (Norris, 1997; Black and Ross, 2015). To do this we first calculate the jump chain matrix, J , of the process from the corresponding generator Q . We then make the set of states corresponding to $Z_3 = 1$ absorbing by setting the rows of J equal to 0 for that set of states, thus there is no probability of leaving these states once entered. Denoting this matrix J' , we solve (Black and Ross, 2015),

$$(I - J')\mathbf{x} = \mathbb{1}_{\{i \in S | z_i = (1, 0, 0, 0, 0)\}}, \quad (17)$$

for \mathbf{x} , where the initial state corresponds to a single exposed individual. Note that \mathbf{x} is a vector of probabilities of visiting or ending in each state, thus to get the hitting probability distribution, ϕ^* , we set the elements of \mathbf{x} corresponding to $Z_3 \neq 1$ equal to zero,

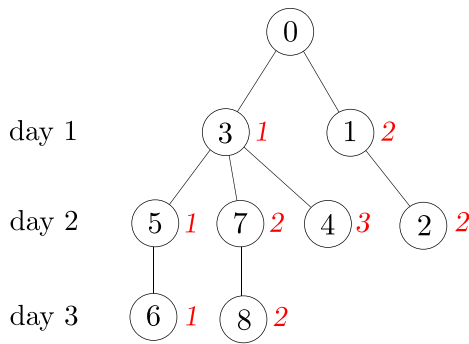
$$\phi^* = \mathbf{x} \odot \mathbb{1}_{\{i \in S | z_{i3} = 1\}}. \quad (18)$$

If a time series has been truncated ($b_j = 1$) then this means that there were no further cases detected after the last observation and this needs to be taken into account in the likelihood calculation, Eq. (8), by multiplying by the probability of observing no further infections after the last detection,

$$L(\mathbf{d}; \theta) = P(Z_3(\infty) = d_n) P(Z_3(t_1) = d_1) \prod_{i=2}^n P(Z_3(t_i) = d_i | Z_3(t_{i-1}) = d_{i-1}). \quad (19)$$

This probability is calculated by integrating forward the pmf of the state until it has converged, denoted ϕ_∞ . Similarly to above, this can be computed by solving a system a linear equations (Black and Ross, 2015),

$$(I - J)\mathbf{x} = \phi(t_n|\mathbf{d}), \quad (20)$$



node (k)	c_k	f_k	\bar{f}_k	children
0				{1, 3}
1	2	0	0	{2}
2	2	0	1	
3	1	0	0	{4, 5, 7}
4	3	1	1	
5	1	0	0	{6}
6	1	0	1	
7	2	0	0	{8}
8	2	0	1	

Fig. 3. Tree corresponding to the example data given in Eq. (21). The nodes of the tree are labelled by the integers $k = 0, \dots, 8$, with 0 denoting the root. The numbers in red are the values of the observed state of the system, c_k , stored by each node, from which the time-series can be recreated. The variables f_k and \bar{f}_k record the number of truncated or continuing time series that are represented in the data that would result from traversing the path from node 0 to node k . For example, $\bar{f}_2 = 1$ indicates that there is 1 time series of the form (2, 2) and this has been truncated, i.e., no further cases were detected after the 2nd day. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where the probability of no further detections is $P(Z_3(\infty) = d_n) = \mathbf{x} \cdot \mathbb{1}_{\{i \in A | z_{i3} = d_n\}}$.

2.4. Efficient data structure for calculation of the likelihood

When we wish to perform inference on case-detection time series from a number of households, the simplest way is to evaluate the product (or the sum when using log-likelihoods) of each individual likelihood as the households are assumed independent. While straightforward, this introduces a large amount of redundancy as the calculations described above must be carried out for each time series and hence computation time scales linearly with the number of households included. As we perform inference in a Bayesian MCMC framework, where the likelihood must be evaluated repeatedly, this is undesirable. In this section we develop a data structure for the efficient representation of this partial case data as well as an algorithm that uses this structure to calculate the likelihood using the minimum number of operations, eliminating all redundancy. In describing this we will assume the data comes from households all of the same size, N . Generalising to a set of sizes simply requires creating a different data structure for each household size in the set.

Consider a new outbreak in a population and assume that by time t there are m households with detected cases, so the data collected from these is a set $\mathbf{D} = \{\mathbf{d}^{(j)}\}_{j=1:m}$ with the truncation status of each series recorded by the vector \mathbf{b} . Thus the total data for performing inference is then defined as $\mathcal{D} = \{\mathbf{D}, \mathbf{b}\}$. For example,

$$\mathcal{D} = \left\{ \{(2, 2), (1, 3), (1, 1, 1), (1, 2, 2), (1, 3)\}, (1, 0, 1, 1, 1) \right\}, \quad (21)$$

where the outbreak in the second household is potentially ongoing as $b_2 = 0$. The key observation is that, for a given household size, \mathcal{D} can be represented as paths through a rooted tree with nodes carrying the data on the cumulative number of detected cases at the end of each day of observation. The minimal tree to represent a given data set is simple to construct and the structure also encodes the minimum number of operations needed to calculate the likelihood of \mathcal{D} . For example, the set of data in Eq. (21) can be represented with the tree shown in Fig. 3.

The nodes of the tree are labelled with the non-negative integers, k , and each node, apart from the root (0) carries three variables: c_k , f_k and \bar{f}_k . The observed state of the system at the end of a given day is recorded by c_k . Thus by starting at the root and traversing downwards, each time series in the set \mathcal{D} can be recreated by accumulating in a list the values of c_j from the visited nodes. The variables f_k and \bar{f}_k count which of all the possible paths

represent continuing and truncated time-series in \mathcal{D} , respectively. For example, the path via the nodes (0, 3, 5, 6) in Fig. 3 recreates the series $(c_3, c_5, c_6) = (1, 1, 1)$ with $\bar{f}_6 = 1$ indicating that there is 1 of these in \mathcal{D} and that it is truncated.

The leaves of the tree always correspond to a time series ($f_k, \bar{f}_k \neq 0$), but so too can other internal nodes. The day of the observation (rescaled within-household time) is given by the depth of a node in the tree, thus nodes at the same level in the tree represent observations on the same day and the maximum depth will correspond to the longest time-series. For this data structure to be more efficient for calculating the likelihood, the populations that are observed must be small but numerous, so there is a large similarity between time series, as is the case here.

2.4.1. Construction of the tree

The construction of a tree representing \mathcal{D} is straightforward as the structure of the tree also provides a way to efficiently search through it. First a root node is created, and then the tree is grown by adding each $\mathbf{d}^{(j)} = (d_i)^{(j)} \in \mathcal{D}$ in turn. The procedure for this is as follows. The algorithm starts at the root of the tree and queries if it has a child node with $c_k = d_1$. If it does, then the algorithm moves to the node. If not, a new child node is created with $c_k = d_1$ and the algorithm then moves to that node. It then repeats this for the remaining elements of $\mathbf{d}^{(j)}$. Once the end of $\mathbf{d}^{(j)}$ is reached, this is recorded by incrementing the value of f_k by 1 if the time-series is continuing, or \bar{f}_k if it is truncated. After this the algorithm returns to the root node and $\mathbf{d}^{(j+1)}$ is added in the same way. The tree shown in Fig. 3 was created in this manner from the data in Eq. (21). This procedure also ensures that the nodes are ordered such that the parent of node k has a label less than k . The total number of nodes is denoted T .

2.4.2. Algorithm for computing the likelihood

Here we detail how this data structure can be used to efficiently calculate the log-likelihood of a given set of observations, \mathcal{D} . The tree is pre-computed before any inference and does not change. The dominant cost in the computation of the likelihood, as detailed in the previous section, is numerically integrating forward the dynamics of the system over a given day. The tree structure encodes the minimum number of these operations that have to be performed to calculate the likelihood, which are only carried out if a node has at least one child. The number of these steps will be the total number of nodes, T , minus the number of leaves in the tree. Hence the computational cost of the likelihood no longer grows linearly with m , the number of household time-series.

To calculate the log-likelihood, each node is associated with a variable $\xi^{(k)}$, that will hold a state vector. We also create two

vectors, ψ and χ , both of length T . These vectors are initialised to be all zeros and also indexed from zero to match the labelling of the nodes of the tree. For the initial step in the algorithm, we calculate $\phi(1)$ from Eq. (13) then set $\xi^{(0)} = \phi(1)$. The algorithm then iterates through the remaining nodes of the tree in either a breadth-first manner or by visiting nodes following the order of their labels (which is determined when the tree is first constructed). For the k 'th node we do the following:

- 1 Set $l = \text{parent}(k)$.
- 2 Calculate $\psi_k = \psi_l + \log \left(\xi^{(l)} \cdot \mathbb{1}_{\{i \in S | z_{i3} = c_k\}} \right)$.
- 3 If either the current node (k) has children or $\bar{f}_k \neq 0$, then set

$$\xi^{(k)} = \frac{\xi^{(l)} \odot \mathbb{1}_{\{i \in S | z_{i3} = c_k\}}}{\xi^{(l)} \cdot \mathbb{1}_{\{i \in S | z_{i3} = c_k\}}} \quad (22)$$
- 4 If $\bar{f}_k \neq 0$, set $\chi_k = \psi_k + \log(\mathbf{x} \cdot \mathbb{1}_{\{i \in A | z_{i3} = c_k\}})$, where \mathbf{x} is the solution of $(I - J)\mathbf{x} = \xi^{(k)}$.
- 5 If the node has children then integrate forward the state vector, $\xi^{(k)}$, by recursively solving the system of equations (12) and storing the result in $\xi^{(k)}$ (overwriting the previous vector).
- 6 Move to the next node.

Once the algorithm has iterated over all nodes the log-likelihood is given by,

$$\log(L(\mathcal{D}; \theta)) = \mathbf{f} \cdot \psi + \bar{\mathbf{f}} \cdot \chi. \quad (23)$$

2.5. Validating the inference methodology

We validated the methodology using data generated from the stochastic household model on which it is based. As well as validation, this allowed us to investigate how the posterior estimates for the parameters improve as more household time-series are added to a given dataset, and to systematically quantify any biases that arise. The methodology was tested on four parameter sets that differ in severity and transmissibility that are chosen to be representative of plausible scenarios of interest to public health agencies (McCaw et al., 2013). The mean latent and infectious periods were fixed for all parameter sets at 2 days and 1.5 days respectively. For severity, the base detection parameter was taken as $p = 0.1$ and 0.5 , for low and high cases respectively. The increased detection parameter was taken as $q = p + 0.4$ representing a constant increase from additional surveillance (McCaw et al., 2013). For transmissibility, the ratio β/γ was taken as 1.2 and 1.4 for low and high scenarios respectively, resulting in a hSAR_3 of 0.49 and 0.54 respectively. The parameters are summarised in Table 2.

An additional complication in generating this test data arises because there is no notion of between household mixing in our stochastic model as the dynamics within each household are assumed independent. The transmission process in the overall population and the detection probabilities will dictate what distribution of household sizes are represented in a random sample and thus it is important to match these when validating the

inference methodology. These distributions were estimated using the microsimulation model (see next section) and are shown in the supplementary material. It was found that the household size distributions only differed significantly between the different severity scenarios and not with transmissibility. As expected, lower severity tends to bias the distribution to larger household sizes because there are more chances for initial detection in a larger household.

With these two distributions as inputs, the test data is generated by first drawing a random sample of 2000 household sizes from the correct distribution and then simulating a within-household epidemic for each, conditioning on at least one detection. This was done using the standard stochastic simulation algorithm (Gillespie, 1976) and then the resulting time series were binned into days. The set of 2000 within-household time series are kept ordered such that first m households ($m \in \{50, 100, 200, 300, 2000\}$) can be selected, which forms the dataset, \mathcal{D} , for inference. The first four of these numbers of households are within the range of data collected in previous studies (House et al., 2012), while 2000, which is an infeasible amount for an FF100 study to actually collect, is to assess convergence with a large amount of data.

2.6. Micro-simulation model for additional validation data

In order to more robustly test our inference methodology, we applied it to data generated by an independent microsimulation model that simulates both disease transmission and surveillance in a population with age and household structure (Geard et al., 2013, 2015). Using a more complex disease transmission model enables us to generate test data that is subject to additional factors that may challenge our inference method. Within the microsimulation, susceptibility to infection varies with age, as observed in recent influenza outbreaks (Opatowski et al., 2011), and mixing between households is parameterised using age-specific patterns of contact. Therefore the subset of households that experience infection, and the order in which they do so is influenced by the number and age of their occupants. Furthermore, households can experience multiple introductions, and the chance of this occurring will vary with both household composition and the current prevalence of disease in the population. Finally, the population is subject to ongoing importation from external sources, which introduces variability in the timing and momentum of an outbreak during its early stages.

We now describe the different components of the model in turn.

2.6.1. Population

The population component of the microsimulation model is calibrated against demographic data to capture age distribution, household size distribution and household composition corresponding to a contemporary Australian population (using parameters and data sources as described in Geard et al., 2013). We use a static population, as demographic changes associated with birth, death, aging and associated changes to households are unlikely to have a substantial impact over the duration of a first few hundred study.

2.6.2. Disease

Disease dynamics were simulated using an SEIR model, with the microsimulation model tracking the current disease state of each member of the population. Durations for the exposed (E) and infectious (I) states were sampled from a Gamma distribution with shape parameter $k=2$ and means of 2 days and 1.5 days respectively. The force of infection acting on each person is a function of current disease prevalence in their household and the broader community. Household structure and contact patterns arise endogenously in the microsimulation model as a result of population demographics. We assume that mixing within the household is frequency dependent and independent of age. Mixing

Table 2

Parameters used to generate the validation data. The parameters β and p are given as low/high pair, the various combination of which generate the four different parameter sets.

Parameter	Values
β	0.8, 0.933
$1/\sigma$	2
$1/\gamma$	1.5
p	0.1, 0.5
q	$p + 0.4$

between households is also frequency dependent and is modelled using an age-specific matrix of contact rates derived from empirical studies (Mosson et al., 2008). Details of the construction of this contact matrix are provided in Geard et al. (2015).

The probability of a susceptible person, s becoming infected in a given time period is given by $1 - e^{-\lambda_s \Delta t}$, where

$$\lambda_s = \psi \left(\sum_{k \in H} \zeta \frac{\tau_h I_k(t)}{(N_H(t) - 1)} + \sum_j \eta_{ij} \frac{\tau_c I_j(t)}{N_j(t)} \right), \quad (24)$$

where ψ is the relative susceptibility of adults compared to children, H is s 's household, N_H is the number of people in H , k is a housemate of s , ζ is the number of contacts per day between s and k (here, we assume $\zeta = 1$; i.e., a susceptible person encounters each of their household members), $I_k = 1$ if k is infectious and 0 otherwise, τ_h is the per contact household transmission coefficient, i is the age of s , η_{ij} is the mean number of contacts in the community per day between people of age i and people of age j , τ_c is the per contact community transmission coefficient, I_j is the number of infectious people of age j and N_j is the total number of people of age j .

The disease state of each individual is updated at discrete intervals of $\Delta t = 3$ h (i.e., 8 time steps per day). We assume that the entire population is susceptible at $t = 0$, with each member having a small probability of being infected from an external source at each time step. We assume that adults (age ≥ 18) have reduced susceptibility compared to children ($\psi = 0.7$) but contribute equally to the force of infection.

2.6.3. Surveillance

Each infected person has a probability p of being detected at the point at which they become infectious (and symptomatic) or otherwise remains undetected. Following the first case detection, enhanced surveillance commences and household members of detected cases are monitored until seven days have passed without any new cases being detected in that household. While under surveillance, people who become infectious have an increased probability $q (\geq p)$ of being detected.

2.6.4. Scenarios

We consider scenarios similar to those for the earlier validation data: low/high severity, and low/high transmissibility. For the high and low transmissibility scenarios, τ_h and τ_c were calibrated such that final attack proportion and secondary household attack proportion were approximately 0.43 and 0.48 respectively. For the high and low severity scenarios, detection probabilities for unmonitored cases were 0.1 and 0.5, increasing to 0.5 and 0.9 respectively for cases occurring among monitored individuals. The microsimulation parameters are summarised in Table 3.

Table 3
Parameters used for the microsimulations.

Parameter	Value
<i>Population</i>	
Population size	100,000
<i>Disease</i>	
Household transmission coefficient (τ_h)	0.97, 1.13
Community transmission coefficient (τ_c)	0.045, 0.055
Mean time in E	2 days
Mean time in I	1.5 days
Updates per day	8
Importation rate	3 cases per week
Adult susceptibility factor (ψ)	0.7
<i>Surveillance</i>	
Case detection probability (p)	0.1, 0.5
Watchlist detection probability (q)	0.5, 0.9

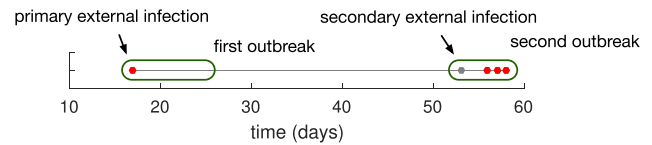


Fig. 4. Example of how microsimulation data is truncated. In this case the second outbreak is removed. Note that we do not remove secondary infections into the household if they occur during the period of time the first outbreak is defined, here 8 days.

2.6.5. Data

The microsimulation outputs the times at which individual cases are detected as well as the times of all other undetected cases so that the entire course of an epidemic can be reconstructed. To make this data suitable for inference using the simpler stochastic model, the raw output from the microsimulation is parsed into detection counts stratified by household and binned into days. In some cases the within-household time series must be truncated as secondary introductions are possible, typically much later in the epidemic, giving rise to further outbreaks. These are removed so that we only consider the first outbreak within a household (although within that first outbreak there may still be multiple introductions). Fig. 4 shows an example of this; note that in this example if the secondary infection occurred within 8 days of the primary infection then the data would have been included as a single outbreak. Finally households of size $N = 1$ or $N > 8$ are removed; single person households contribute no information for inference of within-household parameters and larger households are probably atypical and transmission unlikely to be simply frequency-dependent. They are also, at least in developed countries, rare so are removed for computational efficiency.

3. Results

3.1. Validating the inference methodology

We first validate the methodology using data generated from the same stochastic household model as used for the inference. For each set of data and number of households inference was performed using a Bayesian MCMC algorithm (Gilks et al., 1995) to generate samples from the posterior. Priors were chosen to be uniform for β/γ , $1/\sigma$ and $1/\gamma$, with ranges as detailed in the supplementary material. The prior for p and q was taken to be uniform over the upper triangle defined by $0 > p \geq q \geq 1$. All proposals were independent Gaussian with variances described in the supplementary material. Burn in was 5×10^3 samples and then 5×10^5 samples were taken from the posterior. No thinning was performed on these samples. For each set of posterior samples we used a batch means method to calculate the effective sample size (ESS) for each parameter (Robert and Casella, 1999). Figures summarising these statistics are given in the supplementary material, but all ESS are above 550 and the majority were above 2000.

First we discuss the convergence of these results as data are increased, which is similar across parameter scenarios, then discuss the main differences between the parameter scenarios. Fig. 5 shows how the marginal posteriors for all parameters, from 8 random high transmissibility, high severity data sets, evolve as more households are included in the sample. The true value of the parameters is shown by the black solid line. The 8 posteriors are shown to illustrate the differences, which can be substantial, within and between datasets; only by inspecting the posteriors can we understand how both the mean and variance of the distributions change together. The boxplots at the bottom of each panel in Fig. 5 are calculated from the means of the marginal posteriors. For $m = 50, \dots, 300$ households, the means from 128 posteriors (from independent

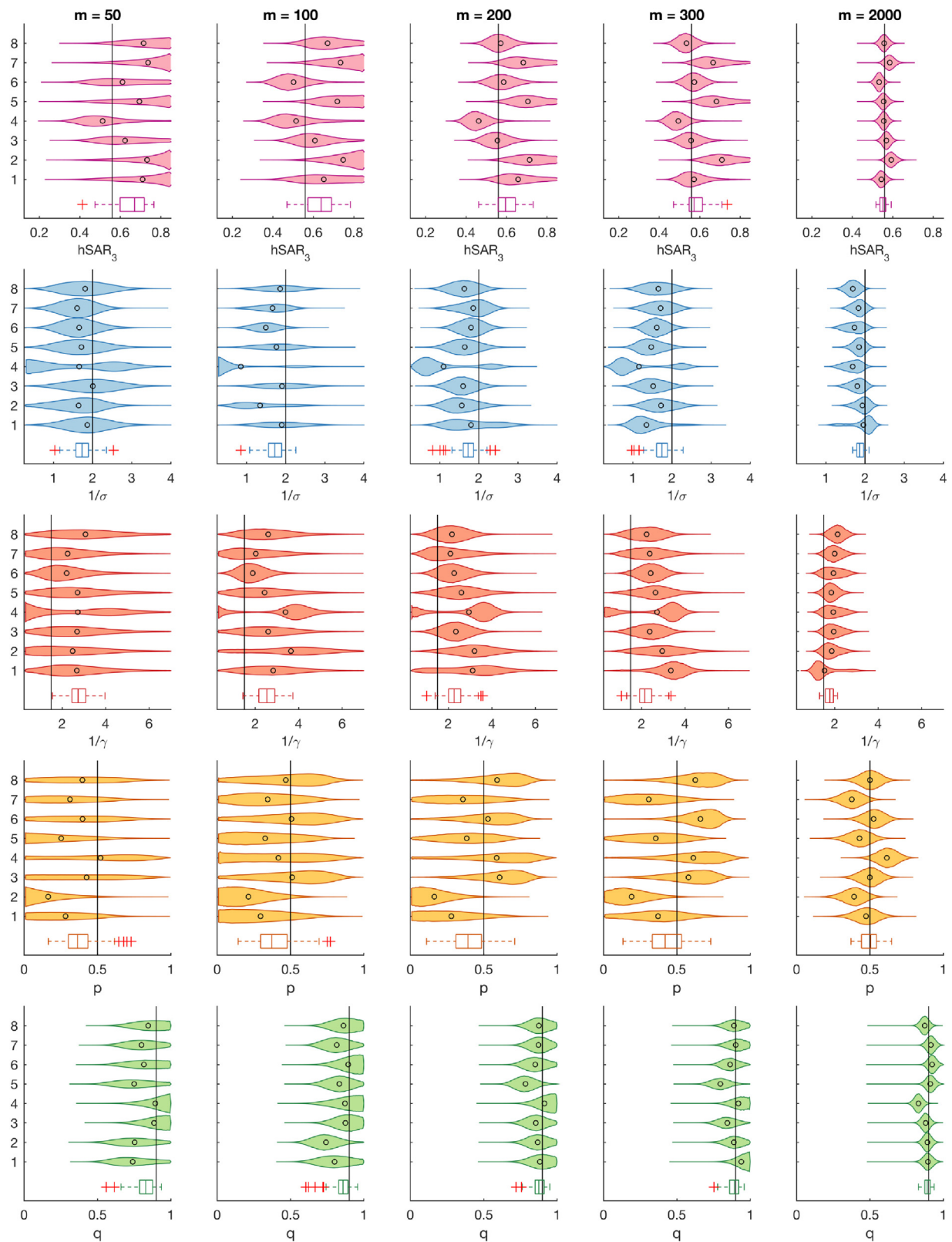


Fig. 5. Marginal posteriors for all parameters from 8 randomly generated, high transmissibility, high severity, data sets using data from the first 50, 100, 200, 300 and 2000 households. The mean of each distribution is shown by the black circle and the true value of the parameter is indicated by the solid line. Note that the scale between each panel is different to aid legibility of the figure. The box plot at the bottom of each panel is derived from the means of 128 (for $m=50, \dots, 300$) and 32 ($m=2000$) marginal posteriors.

data sets) are used to construct the box plots, whilst for $m = 2000$ households, the mean from 32 posteriors are used. Similar plots for the other scenarios are presented in the supplementary material.

In all of these plots we see that as the amount of data is increased the posteriors converge to the true values of the parameters, although the rate of convergence and bias is different between parameters. In general we see that the hSAR and severity, q , converge quickly, and hence are estimated well by 200 households, but the other three parameters have much higher variances. It is also clear from the boxplots of the means of the posteriors, that for smaller amounts of data there tend to be systematic biases in our estimates. The hSAR and mean infectious period are overestimated and the other parameters are slightly underestimated. Note that this is not due to an approximation or assumption in the inference methodology, but highlights the bias and uncertainty to be expected in first few hundred studies due to the small data (Gellman et al., 2013). The priors also have an effect on this, as for small amounts of data the posterior will reflect the prior more than the data. Considering 2000 households in a single sample is probably an unrealistically large amount of data for a real FF100 study to collect, but we have included this to validate the methodology and show that the posteriors are converging correctly as the amount of data increases.

Thus we can see that by $m = 2000$ households, individual posteriors can be biased, but the mean bias is almost completely gone for all but the duration parameters ($1/\gamma$ and $1/\sigma$). These two periods are the hardest to infer accurately from this type of data with posteriors retaining high variance and small amounts of bias even at 2000 households. Inspection of the posteriors for individual datasets suggests that these two parameters are highly correlated. A consistent pattern repeated for all scenarios is that $1/\gamma$ tends to be overestimated and $1/\sigma$ underestimated for finite amounts of data. The differences in posteriors *between* datasets can also be quite pronounced, especially for smaller numbers of households. This is again simply down to the random nature of each outbreak. If more infections or detections than average happen early on then the posteriors will be biased upwards and vice versa.

Now concentrating on the differences between parameter sets, we observe that the largest difference in the inferred posteriors is between the severity scenarios. In the low severity cases there is less data for a given number of households and so inference is less precise (see figures in the supplementary material); both biases and variances are typically larger. Apart from this, the same patterns as noted above also hold for the other scenarios.

3.2. Inference on microsimulation data

In the previous results section we demonstrated that our methodology is able to recover the main parameters of interest from data generated by the same stochastic household model as used for the inference. Real data are unlikely to conform fully to our assumptions and so we now go further and test our inference methodology on data that is generated by the microsimulation model that more accurately reflects true pandemic and social dynamics.

For most of the parameters we wish to estimate from the microsimulation data we know their true values (see Table 3), but due to the age dependent susceptibility, calculating *analytically* an expected value of the hSAR is not possible. We need to estimate this independently of the partial detection process so that we can quantify any biases that result from the partial detection. We can do this by performing inference on the complete household data (detected plus undetected cases) that is available from the microsimulation. Once we have this estimate we perform inference on only the detected data, which is the realistic scenario.

3.2.1. Complete data inference

We perform inference for the underlying epidemiological parameters (β , σ , γ) by using complete case data (using both detected and undetected cases) and assuming $p = q = 1$ in our inference model. As in the validation, we generate 128 datasets for both high and low transmission scenarios and select the first 300 households from each for this inference (this number is somewhat arbitrary, the larger the number the less variance in the estimate). The mean values for the hSAR₃ across the datasets are 0.56 and 0.5, for high and low transmission scenarios respectively. Note that in the microsimulations there is the possibility of external infections after the first case, which violates one of our assumptions in the inference model. This will give rise to slightly more cases so is likely to bias the secondary attack rates higher than if there were no external infections.

3.2.2. Partial detection inference

With the ‘true’ values of the transmissibility estimated, we now perform essentially the same investigation, as done for the validation data, using the microsimulation data. The same MCMC routine and parameters were used as described in Section 3.1 and ESS statistics for the resulting posterior samples are summarised in the supplementary material.

Fig. 6 shows box plots derived from mean estimates of marginal posteriors for an increasing number of household time series for (a) high transmissibility, high severity and (b) high transmissibility, low severity scenarios. As in the validation, each boxplot is constructed from the means of 128 independently generated data sets. Plots similar to Fig. 5, showing marginal posteriors from the first 8 datasets, are shown in the supplementary material.

Both severity scenarios show similar behaviour as seen in the validation results, with convergence of the mean estimates as the number of households time series used is increased. There is once again a similar pattern with respect to transmissibility and the mean latent and infectious periods. A clear difference between the inference on the validation data and the microsimulation data is that both detection probabilities are biased upwards in the microsimulation inference compared with the validation inference. This appears to be true for both severity scenarios.

3.3. Inference on daily incoming data

Another use of our methodology is to look at the behaviour of the posteriors and in particular the transmissibility and severities/detection probabilities p and q , as we increase the data collected during the early stages of an outbreak. When performing inference in this way, in contrast to the previous section, we will have data from potentially ongoing household outbreaks. For these results we set the truncation length as $T_D = 8$ days, that is, a time series is truncated after the number of detected cases has not changed for 8 consecutive days. We also reduce the priors to more closely match an outbreak of influenza. In particular, we set the prior on the hSAR to be approximately uniform over $[0.1, 0.85]$, by choosing an exponential prior on β/γ . The priors on the other parameters are as used for the validation inference and given in the supplementary material.

Instead of creating a contour plot of our inferred marginal posteriors, we split the hSAR₃/ q plane, quantifying transmissibility and severity respectively, into nine regions and shade each region in proportion to the marginal posterior that lies within a given region. This is to reflect a stratification that would be used to inform a pandemic response (Australian Department of Health, 2014). We selected the first 300 households from a microsimulation outbreak with high transmissibility and high severity and the days were calculated when the cumulative number of case detections is first greater than 50, 100, 200, ..., 600 and performed inference on

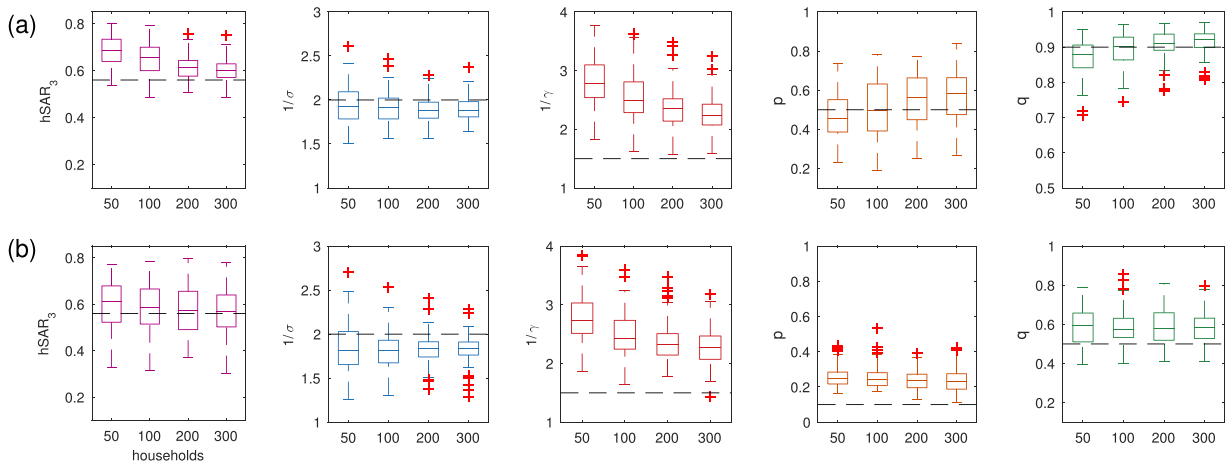


Fig. 6. Convergence of mean estimates from inference on microsimulation data. Each box plot is calculated from marginal mean estimates of posteriors inferred from 128 microsimulations using (a) high transmission/high severity and (b) high transmission/low severity parameters. The ‘true’ values of the parameters as estimated above are shown by dashed lines.

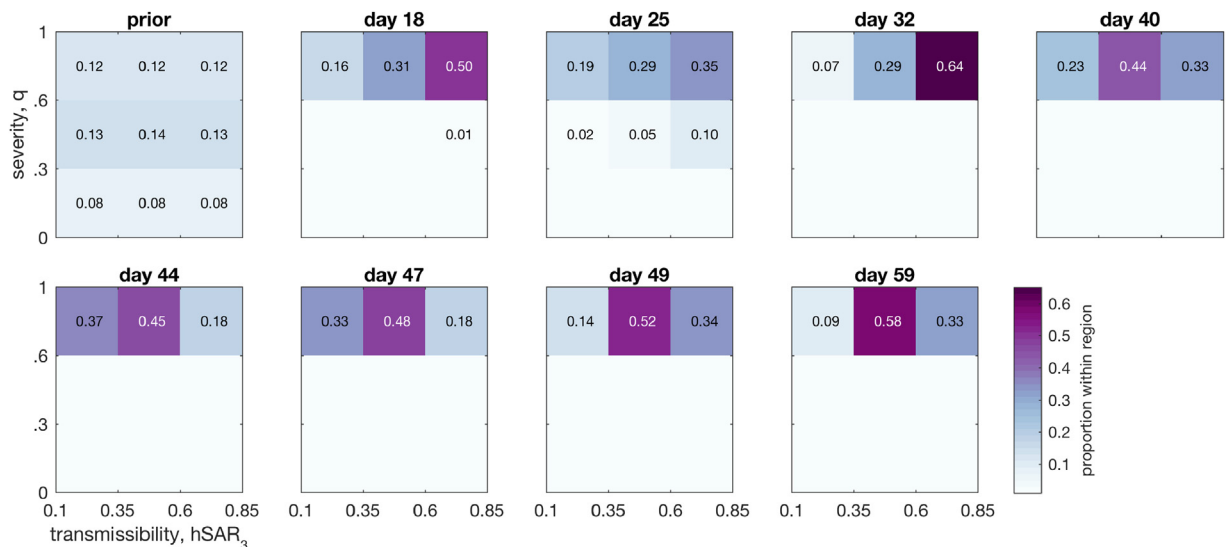


Fig. 7. Assessment of transmissibility and severity of an ongoing epidemic. The $hSAR_3/q$ plane is split into 9 regions and shaded according to the proportion of the posterior samples that fall into a given region. Severity is found quickly, but transmissibility remains variable until much later, being mostly overestimated early on. The true parameters for this outbreak are $hSAR_3 = 0.57$ and $q = 0.9$, corresponding to a high transmissibility/high severity scenario, as determined from inference on full data from the outbreak as described in Section 3.2.1.

the data available up to those days. Fig. 7 shows the evolution of our knowledge over this time frame, as well as the prior. The true values of the parameters are $q = 0.9$ and $hSAR_3 = 0.577 \pm 0.02$. The $hSAR_3$ and error bounds was determined, as in the previous section, from inference on complete data (detected plus undetected cases) from 300 households. Very quickly we are able to characterise the severity of the disease, but the natural variability in the epidemic means it takes considerably longer to characterise the transmissibility. Note that by around day 49 almost all of the 300 households are infected and no further households are added to the data set subsequently. There is still some change in the posterior after day 49, as more data is collected from the households with ongoing infections.

4. Discussion

Early assessment of the likely impact of an emerging influenza pandemic is essential to inform decisions about the appropriate scale of response (Van Kerkhove et al., 2010; Van Kerkhove and Ferguson, 2012; McCaw et al., 2013). FF100 studies are an efficient

and effective means of obtaining data on the time course of infection within individual households. We have developed a method that provides accurate estimates of transmissibility and severity from FF100 data, which have been identified as strong determinants of impact. There are two main novelties to our work. The first is modelling the surveillance process. This allows us to accurately infer detection probabilities and to account for potentially increased surveillance of infected households. It is important to stress that the model we fit is completely mechanistic. The parameter estimates produced may be used subsequently to explore disease dynamics and to evaluate the impact of control strategies. The second novelty of our method is its computational efficiency, which was a goal from the outset of this research. This efficiency not only allows us to perform Bayesian inference, sampling from the full posterior distribution for all parameters, but to repeat this process multiple times over many different realisations. In particular, the inference on 2000 household time series to fully validate the methodology would not be feasible without the tree data structure formulated in Section 2.4. This has allowed us to perform a thorough quantification of the variability and biases inherent to this

method and to test it using an independent model. Such validation is rare in the current literature.

There are a small number of previous studies that are similar to ours (House et al., 2012; Riley et al., 2013, 2015), in working within a transmissibility/severity framework. The most relevant of these is Riley et al. (2015) with the difference being the type of data analysed and the model considered. Riley et al. use data collected from much larger populations (military bases) and fit a deterministic model to each outbreak, deriving a transmissibility/severity estimate for each population. In contrast, we work with data from many households and fit a stochastic model, which is required when dealing with small populations. We also produce a single posterior from data across all households. Our previous work has attempted similar goals, but using only serial interval data instead of full time series and assuming complete detection of cases (Black et al., 2013; Black and Ross, 2013). This paper builds substantially on earlier efforts by conducting inference on full time series with realistic partial detection of cases.

Our analysis has shown that it is possible to estimate both transmissibility and severity accurately from household stratified data. Somewhat surprisingly, it is also possible to estimate the initial severity/detection probability, p , relatively accurately, even for the difficult case when it is very small. The transmissibility within a household is also consistently estimated well. However, the hSAR depends on the product of β and $1/\gamma$, the transmission rate and mean infectious period respectively; individually, β and $1/\gamma$ can be somewhat biased and this trend persists even when full case data are available. With case data available only, the infection process (as illustrated in Fig. 1) does not provide much information on the mean infectious period, $1/\gamma$. This was also found in previous studies using more limited data (Black and Ross, 2013). The mean latent period, $1/\sigma$, also showed some bias, in the direction opposite to that for $1/\gamma$, even with data from 2000 households. The only way to reduce these biases further is to include even more data. It is likely that only some combination of the latent and infectious period parameters can be inferred accurately for smaller data and this is currently being investigated. It is also well known that only the ratio β/γ can be estimated from final size data, but not β and γ individually, or σ at all (Ball et al., 1997; Black and Ross, 2015). Stronger priors would be helpful on the mean infectious and latent periods, making inference on the transmissibility and severity more accurate. It is likely that these will be available early on from other studies where longer chains of infection exist, such as in schools and workplaces (Riley et al., 2013, 2015; Australian Department of Health, 2014).

The inference performs well on both the validation and the microsimulation data, showing the same trends, although the biases are unsurprisingly larger when using the latter. The largest difference uncovered is that inference on microsimulation data leads to the detection probabilities being biased to larger values. From a practical perspective this error is still quite small. For example, a detection probability of 0.93 versus 0.9 would lead us to infer that for 100 observed cases there are approximately 8 unobserved cases when in fact there are 11. From a decision making perspective we only need to bound our estimates to a certain region of transmissibility/severity-space, as we have done when considering the daily updated case data (see Fig. 7).

Both the stochastic household model we use for inference and the microsimulation model make the same strong assumptions about the distribution of the latent and infectious periods as well as the form of transmission (frequency dependent). These assumptions can be relaxed and we might instead wish to estimate these as well by extending the stochastic model, which is straightforward. Often transmission is modelled by a term of the form $\beta/(N-1)^\alpha$, where $-1 \leq \alpha \leq 1$. Other studies have attempted to estimate α (Cauchemez et al., 2004; Kinyanjui et al., 2016) (with some

evidence of $\alpha < 1$), and the same could be done in our model. We have chosen not to do this in this study as our primary goal is to investigate the inherent variability and any biases in the methodology. Doing this requires performing inference on many data sets and hence requires a large amount of computing power. In an actual pandemic situation, with only a single data set, the same amount of computing power could be used to test many different models.

With few exceptions (Cauchemez et al., 2004; O'Dea et al., 2014; Lau et al., 2015), previous work has looked at estimating parameters from single outbreaks (Bettencourt and Ribeiro, 2008; Birrell et al., 2011). For future work it would be interesting to explore the trade-off in parameter estimate accuracy from inference on data collected from an outbreak in a single large population (schools or workplaces) versus outbreaks from a number of smaller ones (households). In smaller populations stochasticity is stronger, but this is potentially offset by independent observations of the same process that are obviously not available with only a single outbreak. Furthermore, the linking of the characterisation of transmissibility and severity to pandemic response strategies, will be an interesting avenue for further research. The only other study similar to ours estimated transmissibility and severity individually for a number of larger populations (Riley et al., 2015). Our method is unsuitable for such large populations and would probably require using approximate methods such as approximate Bayesian computation (ABC) (Toni et al., 2009) or particle MCMC (Andrieu et al., 2010; Golightly and Wilkinson, 2011) to sample from the posterior. We have shown that we can perform inference on a single ongoing outbreak at a daily resolution, but this is relatively costly as it requires running the inference from the beginning of the outbreak for each new day. A true on-line method would be more efficient, but no exact implementations currently exist (Kantas et al., 2015).

There are a number of improvements that can be made to both our model and the inference methodology. We have assumed in our surveillance model that no information can be obtained for any cases before the first detection. This is somewhat pessimistic and the inference could be extended to include these data if they were available, even if only a bound on previous cases could be estimated. The addition of some form of serology could also improve estimates by placing bounds on the numbers of unobserved cases that have occurred in a household. This extra information could then be incorporated into the likelihood with little extra computational cost. Currently we estimate transmissibility only within the household. With estimates of the between-household rate of infection, population level transmissibility can also be estimated (Walker et al., 2016). The microsimulation model that we used to test our inference methodology captures heterogeneities in the population structure (households) and immune status (age-dependent susceptibility). We have however assumed that both transmissibility and severity are uniform across the population. Future work could involve relaxing this assumption to explore scenarios in which transmissibility and severity vary with age.

To our knowledge, this is the first such study to undertake a rigorous analysis of FF100 data and to start to quantify what information is obtainable from it as well as any inherent variability and biases. A key next step will be to look at how much data is needed, and how long it would take to collect, to accurately inform policy. Our analysis of daily incoming data indicates how this could proceed. Surveillance of known contacts beyond members of the immediate household is also a possibility in FF100 studies, but would be substantially more demanding in terms of utilisation of public health capacity. However, this effort may be rewarded by earlier identification of cases, enabling more rapid determination of key epidemic parameters to redirect efforts towards a more targeted response. The trade off between the cost of additional monitoring versus better parameter estimates can be investigated using a combination of microsimulations and our inference

methodology. This combination will allow us to explore many variations on the FF100 study design and therefore to accurately inform the design of real world studies.

Acknowledgements

A.J.B. was supported by an ARC DECRA (DE160100690). J.V.R. was supported by an ARC Future Fellowship (FT130100254). A.J.B. and J.V.R. were also supported by the ARC Centre of Excellence for Mathematical and Statistical Frontiers (CoE ACEMS). N.G. was supported by an ARC DECRA (DE130100660). J.M.M. was supported by an ARC Future Fellowship (FT110100250). J.M. was supported by a Australian Government NHMRC Career Development Award (CDF1061321). A.J.B., N.G., J.M., J.M.M. and J.V.R. were supported by the Australian Government NHMRC Centre for Research Excellence in Policy Relevant Infectious diseases Simulation and Mathematical Modelling (CRE PRISM²).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.epidem.2017.01.004>.

References

- Andrieu, C., Doucet, A., Holenstein, R., 2010. Particle Markov chain Monte Carlo methods (with discussion). *J. R. Stat. Soc. B* 72, 269–342.
- Australian Department of Health, 2014. Ageing Australian Health Management Plan for Pandemic Influenza.
- Baguelin, M., van Hoek, A.J., Jit, M., Flasche, S., White, P.J., Edmunds, W.J., 2010. Vaccination against pandemic influenza A/H1N1v in England: a real-time economic evaluation. *Vaccine* 28, 2370–2384. <http://dx.doi.org/10.1016/j.vaccine.2010.01.002>.
- Ball, F., Mollison, D., Scalia-Tomba, G., 1997. Epidemics with two levels of mixing. *Ann. Appl. Prob.* 7, 46–89.
- Ball, F., 1986. A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Adv. Appl. Prob.* 18, 289–310.
- Bettencourt, L.M.A., Ribeiro, R.M., 2008. Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLOS ONE* 3, e2185.
- Birrell, P.J., et al., 2011. 2011 Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. *Proc. Nat. Acad. Sci. U. S. A.* 108, 18238–18243.
- Black, A.J., Ross, J.V., 2013. Estimating a Markovian epidemic model using household serial interval data from the early phase of an epidemic. *PLoS ONE* 8, e73420.
- Black, A.J., Ross, J.V., 2015. Computation of epidemic final size distributions. *J. Theor. Biol.* 367, 159–165.
- Black, A.J., House, T., Keeling, M.J., Ross, J.V., 2013. Epidemiological consequences of household-based antiviral prophylaxis for pandemic influenza. *J. R. Soc. Interface* 10, 20121019.
- Black, A.J., House, T., Keeling, M.J., Ross, J.V., 2014. The effect of clumped population structure on the variability of spreading dynamics. *J. Theor. Biol.* 359, 45–53.
- 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. *Stat. Med.* 23, 3469–3487. <http://dx.doi.org/10.1002/sim.1912>.
- Cauchemez, S., Donnelly, C.A., Reed, C., Ghani, A.C., Fraser, C., Kent, C.K., et al., 2009. Household transmission of 2009 pandemic influenza A(H1N1) virus in the United States. *N. Engl. J. Med.* 361, 2619–2627.
- Donnelly, C.A., Finelli, L., Cauchemez, S., Olsen, S.J., Doshi, S., Jackson, M.L., Kennedy, E.D., Kamimoto, L., Marchbanks, T.L., Morgan, O.W., et al., 2010. Serial intervals and the temporal distribution of secondary infections within households of 2009 pandemic influenza A (H1N1): implications for influenza control recommendations. *Clin. Infect. Dis.* 52, S123–S130. <http://dx.doi.org/10.1093/cid/ciq028>.
- Geard, N., McCaw, J.M., Dorin, A., Korb, K.B., McVernon, J., 2013. Synthetic population dynamics: a model of household demography. *J. Artif. Soc. Simul.* 16, 8.
- Geard, N., Glass, K., McCaw, J.M., McBryde, E.S., Korb, K.B., Keeling, M.J., McVernon, J., 2015. The effects of demographic change on disease transmission and vaccine impact in a household structured population. *Epidemics* 13, 56–64.
- Gelman, A., et al., 2013. *Bayesian Data Analysis*. Chapman and Hall/CRC.
- Ghani, A., Baguelin, M., Griffin, J., van Hoek, A.J., Cauchemez, S., Donnelly, C., et al., 2009. The early transmission dynamics of H1N1pdm influenza in the United Kingdom. *PLoS Curr.* 1, RRN1130.
- Gibson, G., Renshaw, E., 1998. Estimating parameters in stochastic compartmental models using Markov chain methods. *Math. Med. Biol.* 15, 19–40. <http://dx.doi.org/10.1093/imammb/15.1.19>.
- Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1995. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC.
- Gillespie, D.T., 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* 22, 403–434.
- Golightly, A., Wilkinson, D.J., 2011. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*.
- Health Protection Agency England, 2009. “First Few Hundred” Project Epidemiological Protocols for Comprehensive Assessment of Early Swine Influenza Cases in the United Kingdom.
- House, T., Baguelin, M., van Hoek, A.J., Flasche, S., White, P., Sadique, M.Z., Eames, K., Read, J., Hens, N., Melegaro, A., et al., 2011. Modelling the impact of local reactive school closures on critical care provision during and influenza pandemic. *Proc. R. Soc. B* 278, 2753–2760.
- House, T., Inglis, N., Ross, J.V., Wilson, F., Suleman, S., Edeghere, O., et al., 2012. Estimation of outbreak severity and transmissibility: influenza A(H1N1)pdm09 in households. *BMC Med.* 11, 1–7.
- Jenkinson, G., Goutsias, J., 2012. Numerical integration of the master equation in some models of stochastic epidemiology. *PLoS ONE* 7, e36160.
- Kantas, N., Doucet, A., Singh, S.S., Maciejowski, J., Chopin, N., 2015. On particle methods for parameter estimation in state-space models. *Stat. Sci.* 3, 328–351.
- Kinyanjui, T.M., Cassell, J.A., Guttel, S., Middleton, J., Ross, J.V., House, T., 2016. Scabies in Residential Care Homes: Modelling, Inference and Interventions for Well-Connected Population Sub-Units (submitted for publication).
- Lau, M.S.Y., Cowling, B.J., Cook, A.R., Riley, S., 2015. Inferring influenza dynamics and control in households. *Proc. Natl. Acad. Sci.* 112, 9094–9099.
- McCallum, H., Barlow, N., Hone, J., 2001. How should pathogen transmission be modelled? *Trends Ecol. Evolut.* 16, 295–300. [http://dx.doi.org/10.1016/s0169-5347\(01\)02144-9](http://dx.doi.org/10.1016/s0169-5347(01)02144-9).
- McCaw, J.M., Glass, K., Mercer, G., McVernon, J., 2013. Pandemic controllability: a concept to guide a proportionate and flexible operational response to future influenza pandemics. *J. Public Health* 36, 5–12.
- McKinley, T.J., Ross, J.V., Deardon, R., Cook, A.R., 2014. Simulation-based Bayesian inference for epidemic models. *Comput. Stat. Data Anal.* 71, 434–447.
- McLean, E., et al., 2010. Pandemic (H1N1) 2009 influenza in the UK: clinical and epidemiological findings from the first few hundred (FF100) cases. *Epidemiol. Infect.* 138, 1531–1541.
- 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 Med.* 5, e74.
- Norris, J.R., 1997. *Markov Chains*. Cambridge University Press, Cambridge.
- O’Dea, E.B., Pepin, K.M., Lopman, B.A., Wilke, C.O., 2014. Fitting outbreak models to data from many small outbreaks. *Epidemics* 6, 18–29.
- O’Neill, P.D., Roberts, G.O., 1999. Bayesian inference for partially observed stochastic epidemics. *J. R. Stat. Soc. A* 162, 121–130.
- Opatowski, L., Fraser, C., Griffin, J., de Silva, E., van Kerkhove, M.D., Lyons, E.J., Cauchemez, S., Ferguson, N.M., 2011. Transmission characteristics of the 2009 H1N1 influenza pandemic: comparison of 8 southern hemisphere countries. *PLoS Pathog.* 7.
- Reed, C., Biggersta, M., Finelli, L., Koonin, L.M., Beauvais, D., Uzicanin, A., et al., 2013. Novel framework for assessing epidemiologic effect of influenza epidemics and pandemics. *Emerg. Infect. Dis.* 19, 85–91.
- Riley, P., Ben-Nun, M., Armenta, R., Linker, J.A., Eick, A.A., Sanchez, J.L., et al., 2013. Multiple estimates of transmissibility for the 2009 influenza pandemic based on influenza-like-illness data from small US military populations. *PLoS ONE* 9, e1003064.
- Riley, P., Ben-Nun, M., Linker, J.A., Cost, A.A., Sanchez, J.L., George, D., Bacon, D.P., Riley, S., 2015. Early characterization of the severity and transmissibility of pandemic influenza using clinical episode data from multiple populations. *PLoS Comput. Biol.* 11, e1004392.
- Robert, C.P., Casella, G., 1999. *Monte Carlo Statistical Methods*. Springer, New York.
- Sarkka, S., 2013. *Bayesian Filtering and Smoothing*. Cambridge University Press.
- Sunkara, V., 2009. The chemical master equation with respect to reaction counts. In: *Proc. 18th World IMACS/MODSIM Congress*, pp. 703–707.
- Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M., 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* 6, 187–202.
- van Gageldonk-Lafeber, A.B., van der Sande, M.A., Meijer, A., Friesema, I.H., Donker, G.A., Reimerink, J., et al., 2012. Utility of the first few 100 approach during the 2009 influenza A(H1N1) pandemic in the Netherlands. *Antimicrob. Resist. Infect. Control* 1, 30.
- van Kampen, N.G., 1992. *Stochastic processes in physics and chemistry*. Elsevier, Amsterdam.
- Van Kerkhove, M.D., Ferguson, N.M., 2012. Epidemic and intervention modelling – a scientific rationale for policy decisions? Lessons from the 2009 influenza pandemic. *Bull. World Health Organ.* 90, 306–310.
- Van Kerkhove, M.D., Asikainen, T., Becker, N.G., Bjorge, S., Desenclos, J., dos Santos, T., et al., 2010. Studies needed to address public health challenges of the 2009 H1N1 influenza pandemic: insights from modeling. *PLoS Med.* 7, e1000275.
- Walker, J.N., Ross, J.V., Black, A.J., 2016. Inference of Epidemiological Parameters from Household Stratified Data. *arXiv: 1609.09170* (submitted for publication).