Epidemics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

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 xxx

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.

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).

http://dx.doi.org/10.1016/j.epidem.2017.01.004

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.

Please cite this article in press as: Black, A.J., et al., Characterising pandemic severity and transmissibility from data collected during first few hundred studies. Epidemics (2017), http://dx.doi.org/10.1016/j.epidem.2017.01.004