Epidemics 16 (2016) 17–26

Contents lists available at ScienceDirect

Epidemics

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

Information content of household-stratified epidemics

T.M. Kinyanjui[a,*], L. Pellis[b], T. House[a,b]

[a] School of Mathematics, University of Manchester, Manchester M13 9PL, United Kingdom
[b] Warwick Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom

ARTICLE INFO

Article history:
Received 19 August 2015
Received in revised form 15 February 2016
Accepted 25 March 2016
Available online 26 March 2016

Keywords:
Households model
Study design
Data collection
Parameter estimation

ABSTRACT

Household structure is a key driver of many infectious diseases, as well as a natural target for interventions such as vaccination programs. Many theoretical and conceptual advances on household-stratified epidemic models are relatively recent, but have successfully managed to increase the applicability of such models to practical problems. To be of maximum realism and hence benefit, they require parameterisation from epidemiological data, and while household-stratified final size data has been the traditional source, increasingly time-series infection data from households are becoming available. This paper is concerned with the design of studies aimed at collecting time-series epidemic data in order to maximize the amount of information available to calibrate household models. A design decision involves a trade-off between the number of households to enrol and the sampling frequency. Two commonly used epidemiological study designs are considered: cross-sectional, where different households are sampled at every time point, and cohort, where the same households are followed over the course of the study period. The search for an optimal design uses Bayesian computationally intensive methods to explore the joint parameter-design space combined with the Shannon entropy of the posteriors to estimate the amount of information in each design. For the cross-sectional design, the amount of information increases with the sampling intensity, i.e., the designs with the highest number of time points have the most information. On the other hand, the cohort design often exhibits a trade-off between the number of households sampled and the intensity of follow-up. Our results broadly support the choices made in existing epidemiological data collection studies. Prospective problem-specific use of our computational methods can bring significant benefits in guiding future study designs.

1. Introduction

Mathematical models have been identified as important tools in the description of the transmission of infections as well as the evaluation of control strategies (Keeling and Rohani, 2007; Anderson and May, 1991). Early infection models frequently assumed that the population mixed homogeneously with frequency- or density-dependent transmission (Anderson and May, 1991). The homogeneous-mixing assumption can be extended relatively straightforwardly to allow for host heterogeneities such as stratification by age (Anderson and May, 1982; Schenzle, 1984; Keeling and Rohani, 2007). Further extensions involve dividing the population into activity-based risk groups (Hadeler and Castillo-Chavez, 1995; Sutton et al., 2012) or households (Becker and Dietz, 1995; Ball and Neal, 2002; House and Keeling, 2009).

For a number of infections requiring close contacts, transmission within the household (generally defined as a group of individuals sharing living arrangements) has been identified as an important component of spread (Munywoki and Koech, 2013; Cauchemez et al., 2014) due to the greater intimacy and the stable nature of the contacts compared to contacts outside the households (Longini et al., 1982; Read et al., 2008). This has led to the development of household driven dynamic models for the exploration of targeted vaccination programmes (Ball et al., 1997; Becker and Starczak, 1997; House and Keeling, 2009; Poletti et al., 2015). Following their development and more recent usage, these models require parameterization by fitting to household-stratified infection data, typically on final outcomes (O'Neill et al., 2000; Demiris and O'Neill, 2005; Neal, 2012). Advances in laboratory techniques mean that more detailed, temporal, data have increasingly become available (Munywoki and Koech, 2013; Cowling et al., 2009; Horby et al., 2012; Hayward et al., 2014) although these remain costly and time consuming to collect, motivating the question of whether the design of these studies can be optimised.

In order to design a study, choices have to be made on overall protocol, the number of participants, duration, the number of time

* Corresponding author. Tel.: +44 7879219771.

http://dx.doi.org/10.1016/j.epidem.2016.03.002

points to sample, the sensitivity and specificity of tests, and many other questions – all of which should be guided by both knowledge of the system to be measured and resource constraints. This paper addresses the question of designing studies to collect household epidemic data in order to maximize the information available to calibrate the parameters of a household stratified epidemic model given a fixed budget. Household stratified data collection usually involves enrolling households and prospectively following them up to collect samples for pathogen identification. In designing these studies, two main decisions need to be made, with the first being the number of households to enroll and the second being the frequency of data collection or the number of times to collect samples from individuals.

Previous work done by Klick et al. (2012, 2014) evaluated study designs that make most cost-effective use of resources for accurately and robustly estimating the secondary attack proportion (SAP) from a set of households in a transmission study and for maximising statistical power. These studies were carried out within the framework of classical optimal design of experiments and were not concerned with estimation of the parameters of a fully mechanistic, temporal, non-linear epidemic model, instead focusing on careful estimation of a static proportion of secondary infections. On the other hand, work by Cook et al. (2008) considered optimisation of the exact set of time points at which the SI epidemic model is observed, but restricted to one population rather than a population of households.

Here, we provide for the first time a systematic method to optimise information content of household-stratified studies of infection over time at fixed cost, which involves the evaluation of an optimal trade-off between the sample size (number of households enrolled) and the intensity of follow-up (number of time points at which we assume all households are observed). Since the models involved do not have simple likelihood functions, we adopt a Bayesian experimental design framework which enables, amongst other things, the use of a computationally intensive Markov chain Monte Carlo (MCMC) methodology to deal with arbitrary likelihoods. Lindley (1970, pp. 19–20) presents a decision theoretic approach to experimental design, arguing that a good way to design experiments is to specify a utility function which should reflect the purpose of the experiment. Since the main goal of the current work involves making inference on model parameters, we have used a utility function based on Shannon information (Shannon, 1948), a popular choice in Bayesian optimal experimental design that captures many of our intuitions about information (Chaloner and Verdinelli, 1995) and which we discuss in more depth in the Methods section below. Our design choice is, overall, regarded as a decision problem selecting the design that maximises the expected utility.

Competing study designs will be evaluated under two protocols: (1) longitudinal/cross-sectional and (2) cohort. Under the cross-sectional model, the assumption is that the households are randomly selected at every time-point the samples need to be taken, while the cohort model assumes that the same households are followed and sampled throughout the study period. We note that the estimates of information content we provide cannot be used to compare these two protocols. In practice, however, we expect that considerations such as gaining informed consent, recruitment and retention of participants and other practical considerations will take precedence in determining the overall study protocol. This may in fact lead to a hybrid design where new households are chosen at each time-point from within a larger pre-specified grouping – our cross-sectional design emerges from such a hybrid in the limit of a large grouping, and the cohort in the limit of a small grouping – with an example of such an approach being the virological confirmation of selected www.flusurvey.org.uk participants.

In the next sections, we describe the household model, the optimal design formulation including the utility function, the results and a general discussion.

2. Materials and methods

2.1. The household model

We consider the realistic scenario in which the number of households in the population is large, so the overall epidemic is well approximated by its deterministic limit (Ball, 1999; House and Keeling, 2008; Ball and Neal, 2002). We also assume that the number of households as a whole is much larger than the number of experimentally sampled households, so that the observed state of the sampled households bears negligible impact on the epidemic dynamics in the rest of the population.

We will also consider a pathogen for which individuals develop permanent immunity following infection, leading to an SIR compartmental model with S, I and R representing the proportion of the population that is in the susceptible, infected and removed (immune) classes respectively. The deterministic dynamics of this model in the absence of demography have been well studied (Anderson and May, 1991) and correspond to the special case of our general formalism where all households are of size 1 (or where within-household transmission does not occur):

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI, \quad \frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - \gamma I, \quad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I. \tag{1}$$

Here β and γ represent the global transmission rate and the rate of recovery from infection respectively.

To model household-stratified transmission, individuals are assumed to retain their global contacts within the population and also experience an extra force of infection at a rate τ per infectious member within the household. The model is therefore composed of two transmission rates: one representing transmission between susceptible-infected pairs the same household, τ, and the other representing transmission between general members of the community, β. The proportion of households with s susceptibles, i infectives and r recovered individuals at time t is represented by $P_{s,i,r}(t)$, and the proportion of the overall population that is infective is

$$I(t) = \frac{\sum_{s,i,r} iP_{s,i,r}(t)}{\sum_{s,i,r}(s+i+r)P_{s,i,r}(t)}. \tag{2}$$

The complete dynamics are modelled by considering all the possible household infection configurations with the full dynamics determined by the 3 processes visualised in Fig. 1A–C: within household transmission (rate τ); random transmission between individuals in the population (rate β); and recovery from infection (rate γ). The dynamics are therefore described by a set of ordinary differential equations (ODEs)

$$\begin{aligned}\frac{\mathrm{d}P_{s,i,r}}{\mathrm{d}t} &= \gamma\left[-iP_{s,i,r} + (i+1)P_{s,i+1,r-1}\right] \\ &\quad + \tau\left[-siP_{s,i,r} + (s+1)(i-1)P_{s+1,i-1,r}\right] \\ &\quad + \beta I(t)\left[-sP_{s,i,r} + (s+1)P_{s+1,i-1,r}\right].\end{aligned} \tag{3}$$

A rigorous derivation of Eq. (3) can be found in the literature (Ball, 1999; House and Keeling, 2009, 2008). This system does not have a solution in terms of elementary analytic functions, but can be integrated numerically. This requires some care since there are multiple time scales in the system – intuitively, the timescales associated with the progression of the epidemic in the general population, and the (shorter) timescales associated with the progression of a within-household epidemic – that make the system numerically