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

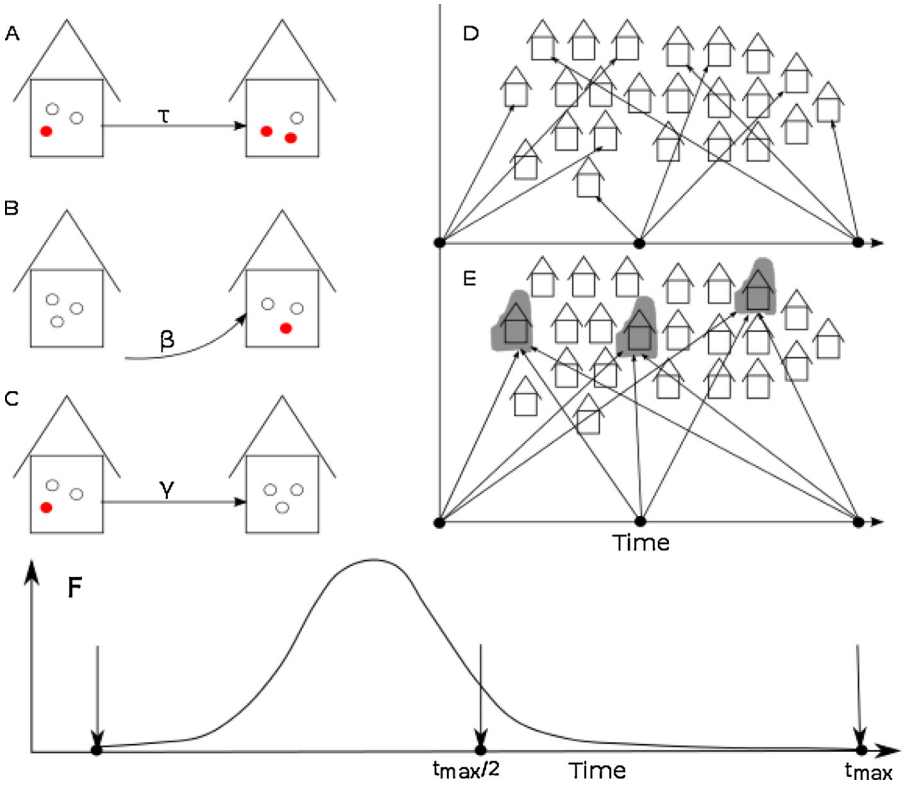

Fig. 1. Graphical representation of model structure. (A) shows the household transmission; (B) shows community transmission and (C) shows recovery from infection, which are all the possible events. The rate of occurrence are τ, β and γ respectively, which are the parameters to be inferred. (D) and (E) show the cross-sectional and cohort study designs, respectively, with 3 households enrolled and sampled 3 times over the course of the study period. (F) shows the 3 sample times evenly distributed over the course of an epidemic. Note that in all the simulations we assume that (i) the initial condition of the system at time zero is known and (ii) the final time point, $t_{\max}$, taken at the end of the epidemic, is always known and part of the sampling scheme.

'stiff'. Since our methods require both Monte Carlo and numerical errors to be controlled, implying a trade-off between speed and accuracy of numerical ODE solver, we used the low-tolerance implicit algorithm `ode23s`, a built-in method in MATLAB based on the order-2 Rosenbrock formula (Mathworks Inc, 2014).

2.2. Framework for optimal study design

We consider our study design problem in a Bayesian framework. Suppose we already have a model characterised by a parameter vector $\boldsymbol{\theta}$ (in our case, $\boldsymbol{\theta}=(\tau, \beta, \gamma)$). We let our prior knowledge be captured by a probability distribution function $\pi(\boldsymbol{\theta})$. Let $\mathcal{E}$ represent the set of study designs such that a generic design $E \in \mathcal{E}$ involves the collection of data D with associated likelihood function $L_E(D|\boldsymbol{\theta})$. The posterior distribution function over parameters is then given by Bayes' theorem,

$$f_E(\boldsymbol{\theta}|D) = \frac{L_E(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L_E(D|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})\mathrm{d}\boldsymbol{\vartheta}}. \tag{4}$$

We then seek a measure of the information content of the posterior distribution. A suitable choice is the Shannon information (Shannon, 1948), which takes larger values for more 'concentrated' posteriors, i.e. with more probability mass in a smaller region of parameter space, and at the same time satisfies a property of additivity, so that the information content in the simplest scenarios will be approximately proportional to the number of observations.

Explicitly, the amount of information per observed data point for a given study design (E) and dataset (D) is

$$H_E(D) = -\frac{1}{\mathcal{N}}\int f_E(\boldsymbol{\theta}|D)\log(f_E(\boldsymbol{\theta}|D))\mathrm{d}\boldsymbol{\theta}, \tag{5}$$

where $\mathcal{N}$ is the number of observations made in the study design. We use this normalisation convention on the reasonable assumption that each observation has similar cost, in order to make a direct comparison between studies of different overall expense.

Fig. 2 shows this in practice, with marginal posteriors on each of the epidemiological parameters becoming 'narrower' as observations are added, at comparable levels of information per observation.

2.3. Likelihood function

In the collection of household stratified epidemic data, we will consider two study protocols. In the first, households to be enrolled are randomly chosen at each time point (a cross-sectional design, shown in Fig. 1D) and in the second, households are randomly chosen at the beginning of the study and prospectively followed for the duration of the study i.e. until the end of the epidemic (a cohort design, shown in Fig. 1E). The parameters to be inferred are the within-household transmission, τ, community transmission, β, and rate of recovery from infection, γ, which are shown in Fig. 1A, B, C respectively.

We will also make the simplifying assumption that the population is made up of households each with a fixed number of members, n; this means that we only need to keep track of the proportion of households with s susceptibles and i infectives, $P_{s,i}$, since the number of recovered individuals will simply be $r = n - s - i$.

2.3.1. Cross-sectional design

In this section, we show the calculation of the full likelihood of observing the data given the model parameters assuming that new households are randomly chosen at each sampling time point. Let

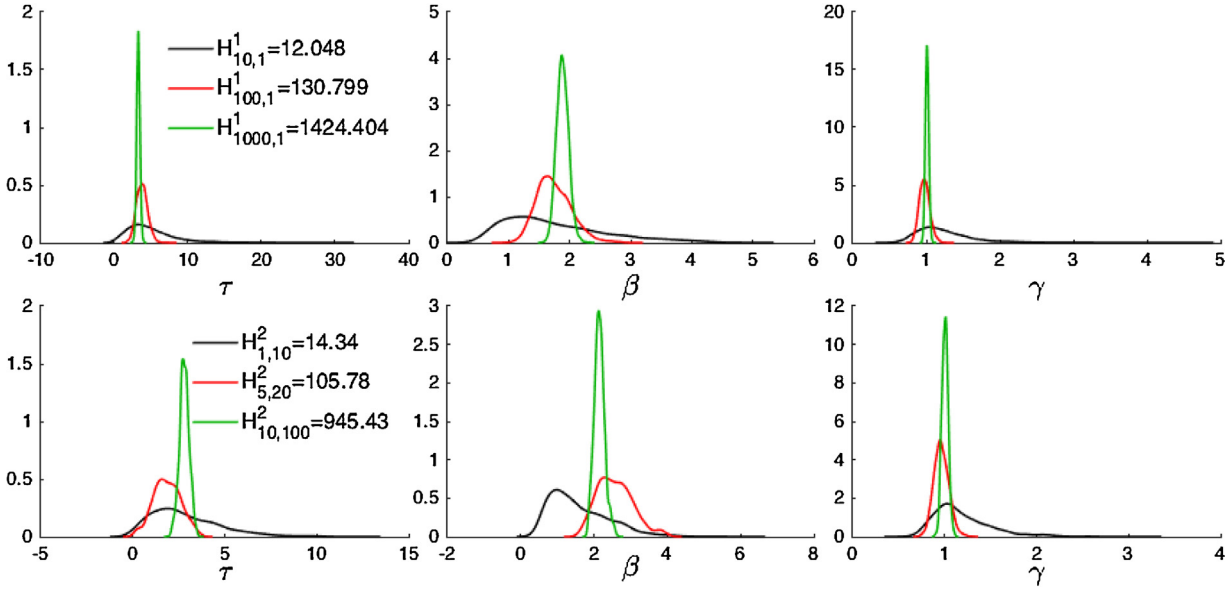

Fig. 2. Concentration of the marginal posteriors with increasing overall information but comparable information-per-observation between the two designs (i.e. cross-sectional and cohort) at optimal sampling intensities for (left) within-household transmission τ (middle) between-household transmission β (right) recovery rate γ; (black) 10 (red) 100 and (green) 1000 observations; (top) cross-sectional (bottom) cohort protocols.

$\mathcal{E}^1 = \{E_d^1\}$ represent the set of all cross-sectional study designs we consider (denoted throughout by the index 1). A generic member of this set will take the form $\left(\mathcal{T}^1, \mathcal{H}^1\right)$ where $\mathcal{T}^1$ is a set of time points such that $T = |\mathcal{T}^1|$ ($|\cdot|$ denotes the number of elements in set $\cdot$) and $\mathcal{H}^1 = (\mathcal{H}(1), \mathcal{H}(2), \ldots, \mathcal{H}(T))$ is a set of sets of households sampled uniformly at random at each time point.

We will make the simplifying, but realistic, assumptions that: (i) the epidemic starts at time 0, (ii) the time points are evenly spaced throughout the study period, meaning $\mathcal{T}^1 = \left(\frac{t_{\max}}{T}, \ldots, t_{\max}\right)$; and (iii) $\forall a \in \mathcal{T}^1$, $|\mathcal{H}(a)| = K$. Therefore $E^1_{(T,K)}$ can be used to represent a cross-sectional study design with T time points and K households. The data for such an experiment takes the form $D = \{Z_{s,i}(a)\}$ where $Z_{s,i}(a)$ is an integer between 0 and K representing the number of households observed in the sample with s susceptibles and i infectives at time point $a \in \mathcal{T}^1$.

The likelihood of observing the data D given the parameter set $\boldsymbol{\theta}$ is therefore given by the product form

$$L_{E^1_{(T,K)}}(D|\boldsymbol{\theta}) = \prod_{a\in\mathcal{T}^1} \Pr\left[Z_{s,i}(a) \mid \boldsymbol{\theta}\right]. \tag{6}$$

The probabilities for each household observation at a given time are then given by the multinomial probability mass function (pmf)

$$\Pr\left[Z_{s,i}(a) \mid \boldsymbol{\theta}\right] = \frac{K!}{\prod_{s,i} Z_{s,i}(a)!} \prod_{s,i} P_{s,i}\left(a;\boldsymbol{\theta}\right)^{Z_{s,i}(a)}. \tag{7}$$

$P_{s,i}\left(a;\boldsymbol{\theta}\right)$ represents the probability of observing a household with s susceptibles and i infecteds given the parameter set $\boldsymbol{\theta} = (\tau, \beta, \gamma)$ at time point a which is obtained by solving Eq. (3) subject to initial conditions

$$P_{s,i}(0;\boldsymbol{\theta}) = \begin{cases} 1-\epsilon & \text{if } s = n \text{ and } i = 0, \\ \epsilon & \text{if } s = n-1 \text{ and } i = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

2.3.2. Cohort design

In this section, we show the calculations of the full likelihood for the second type of design which follows the same households over the course of an entire epidemic. Analogously to the cross-sectional case, let $\mathcal{E}^2 = \{E_d^2\}$ represent the set of all cohort study designs. Then $E^2 = \left(\mathcal{T}^2, \mathcal{H}^2\right)$ where $\mathcal{T}^2 = \left(\frac{t_{\max}}{T}, \ldots, t_{\max}\right)$ is a set of time points such that $T = |\mathcal{T}^2|$ and $\mathcal{H}^2 = (\mathcal{H}(1), \mathcal{H}(2), \ldots, \mathcal{H}(T))$ is a set of sets of households sampled at each time point. Since in the cohort design the aim is to follow the same households over the entire time period, we assume that, in addition to $\forall a \in \mathcal{T}^2$, $|\mathcal{H}(a)| = K$ where K is the number of households enrolled, in contrast to the cross-sectional case $\forall a, b \in \mathcal{T}^2$, $\mathcal{H}(a) = \mathcal{H}(b)$. A generic household cohort study with T time points and K households will therefore be represented by $E^2_{(T,K)}$.

The data for such an experiment takes the form $D = \{Z^k_{s,i}(a)\}$ where $Z^k_{s,i}(a)$ is an indicator variable taking the value 1 if the k-th household in the cohort is observed to have s susceptibles and i infectives at time point $a \in \mathcal{T}^2$, and the value 0 otherwise. The likelihood of observing the data D given the parameter set $\boldsymbol{\theta}$ is also given by the product form

$$L_{E^2_{(T,K)}}(D|\boldsymbol{\theta}) = \prod_{a\in\mathcal{T}^2} \Pr\left[Z^k_{s,i}(a) \mid \boldsymbol{\theta}\right]. \tag{9}$$

Note that, strictly speaking, the probability on the right hand size of (9) should be conditioned on previous time-points as well; we will deal with this further below. The probabilities for each household observation at a given time are then obtained from the multinomial probability mass function

$$\Pr\left[Z^k_{s,i}(a) \mid \boldsymbol{\theta}\right] = \prod_{s,i} Q^k_{s,i}(a;\boldsymbol{\theta})^{Z^k_{s,i}(a)}. \tag{10}$$

Here $Q^k_{s,i}(a;\boldsymbol{\theta})$ is the probability of the k-th household in the cohort being in the configuration with s susceptibles and i infectives; as before, we have the initial condition

$$Q^k_{s,i}(0;\boldsymbol{\theta}) = \begin{cases} 1-\epsilon & \text{if } s = n \text{ and } i = 0, \\ \epsilon & \text{if } s = n-1 \text{ and } i = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

What is fundamentally different from the cross-sectional case is that we need to keep track of the probabilities of each household in the cohort being in a given state, so that, between observations,

the probabilities obey the ODE system

$$\frac{dQ^k_{s,i}}{dt} = \gamma\left[-iQ^k_{s,i} + (i+1)Q^k_{s,i+1}\right] + \tau\left[-siQ^k_{s,i} + (s+1)(i-1)Q^k_{s+1,i-1}\right] + \beta I(t)\left[-sQ^k_{s,i} + (s+1)Q^k_{s+1,i-1}\right]. \quad (12)$$

Here $I(t)$ is defined as in (2) and we assume that the cohort is a negligible fraction of the population so that the dynamics in the rest of the population are described by the proportions $P_{s,i}(t)$ obeying (3). At each observation time point we evaluate the functions (10) and then set the probabilities to the observed values

$$Q^k_{s,i}(a) = Z^k_{s,i}(a) \quad (13)$$

as new initial conditions for integrating (12) over the next time interval $[t_a, t_{a+1}]$ and derive $Q^k_{s,i}(a+1)$.

For both the cross-sectional and cohort designs, it is important to note that we assume that the initial condition of the system at time zero is known and that the final time point occurs after the epidemic has finished and is always recorded as part of the sampling scheme as shown in Fig. 1F. A visualisation of the structure of simulated cohort data is given as Supplementary Fig. S1.

2.4. Parameter choices

We generate data D by simulating the models as described starting from 'true' parameters $\boldsymbol{\theta}^* = (\beta^*, \tau^*, \gamma^*)$ and with a given value of household size n set to a typical size for the population in question. We consider two parameter sets: (1) 'RSV-like', which are reasonable values for one seasonal epidemic of Respiratory Syncytial Virus (RSV) in Kenya: $\beta^* = 2$, $\tau^* = 3$, $\gamma^* = 1$, $n = 10$. (2) 'Flu-like', which are reasonable values for Influenza in England: $\beta^* = 1.12$, $\tau^* = 0.96$, $\gamma^* = 0.8$, $n = 4$. These are not intended to be precision estimates, but rather to lie within the range of reasonable values suggested by previous work on RSV and influenza (Frank et al., 1981; Glezen et al., 1986; Carrat et al., 2008; Baguelin et al., 2010; Okiro et al., 2010; House et al., 2012). Although RSV displays SIRS-like dynamics in the long term (i.e. individuals will become re-infected several times in their lifetime) we have modelled it within an SIR framework as very few, if any, re-infections occur within the same seasonal epidemic since the time scale for the loss of acquired immunity is longer than the duration of the epidemic. We note that there may be other infections that offer a more contrasted scenario in terms of transmission potential. However, presence of pre-existing immunity either naturally or vaccine induced will require that we consider a more complicated epidemiological model. Depending on the nature of the study objective, this can be modified so as to reflect a more realistic history of natural infection.

For all the simulations, ϵ, which is the proportion of households in the population at the beginning of the study with one infected and all the other household members susceptible is taken to be 10^{-3}; this quantity is not of biological interest and therefore and we assume that it is known (in the same way that we do not count time 0 as an observation). This is consistent with a naive infection being introduced in the population (e.g. a novel strain of influenza) or an infection whose immunity wanes over time (e.g. RSV) and is chosen to be small enough not to deplete the susceptible population significantly, but to be large enough that we need not consider stochastic effects at the population level (Keeling and Rohani, 2007).

To examine whether the results are robust to changes in the model parameters, we have, for the designs with 100 data points, systematically explored the values of within (τ) and between household transmission (β) and the number of individuals in a household (N) as shown in Table 1. Each of the letters in the first column refers to the corresponding subplot in Figs. 6 and 7. For each parameter set explored, we generated 20 replicates and plotted the resulting information per datapoint.

Table 1
List of parameter combinations explored in Figs. 6 and 7.

Subplot	Household transmission (τ)	Community transmission (β)	Household size (n)
A	0.96	1.12	4
B	2	1.12	4
C	0.5	1.12	4
D	0.96	3	4
E	0.96	0.6	4
F	0.96	1.12	2
G	0.96	1.12	8

2.5. MCMC methodology

To obtain samples from the model posterior distribution, we use Markov Chain Monte Carlo (MCMC) (Robert and Casella, 2010) with Random-Walk Metropolis Hastings sampling, independent Gaussian proposal densities tuned by hand and a starting point at the 'true' parameters $\beta^*, \tau^*, \gamma^*$. Burn-in time was 10^3 and samples were thinned by a factor of 10. Mixing was assessed via trace plots and the total number of samples visualised is 10^3.

The output of this algorithm for each scenario is a set $\{\boldsymbol{\theta}_j\}_{j=1}^M$ of samples from the joint posterior; we therefore estimate the information content per observation of a dataset under study design $E_{(T,K)}$ from (5) as

$$H_{E_{(T,K)}}(D) \approx \frac{-1}{M \times T \times K}\sum_{j=1}^{M}\log(f(\boldsymbol{\theta}_j|D)). \quad (14)$$

In (14), $f(\boldsymbol{\theta}_j|D)$ is the posterior density, which is proportional to the likelihood, with the constant of proportionality depending very sensitively on the priors, which in our case are flat and uninformative. The higher the value of H, the more the information in the designs selected and vice versa. For the RSV-like scenarios, we visualise the full posterior and report H for the first simulated dataset and then generate additional simulated datasets for which we calculate H and plot the region where the 95% credible interval (CI) of the information lies for all the designs. For the flu-like scenario, we do not show the full posteriors but report H and variability through replicates in the same way, as well as introducing additional variability through an additional set of replicates in which simulation parameters are picked from a normal distribution with mean equal to the initial true value and variance equals to 0.2.

It is worth noting that our approach is designed to be capable of adaptation to a more fully Bayesian approach or even use within a frequentist framework. For the former, rather than use uninformative priors as we have done, informative priors could be used and the information gain (i.e. difference in Shannon entropy between prior and posterior) calculated. For the latter, MCMC should be viewed as a versatile method of likelihood exploration for a complex model.

3. Results

3.1. RSV-like parameters

Each of the subplots in Fig. 3 (and Supplementary Figs. S2 and S3) shows the samples from the joint posterior densities of the three parameters β, τ, and γ, and also the projections of these onto the x–y, y–z and x–z planes respectively. The points are shaded according to their log-likelihood with the highest values in yellow and the

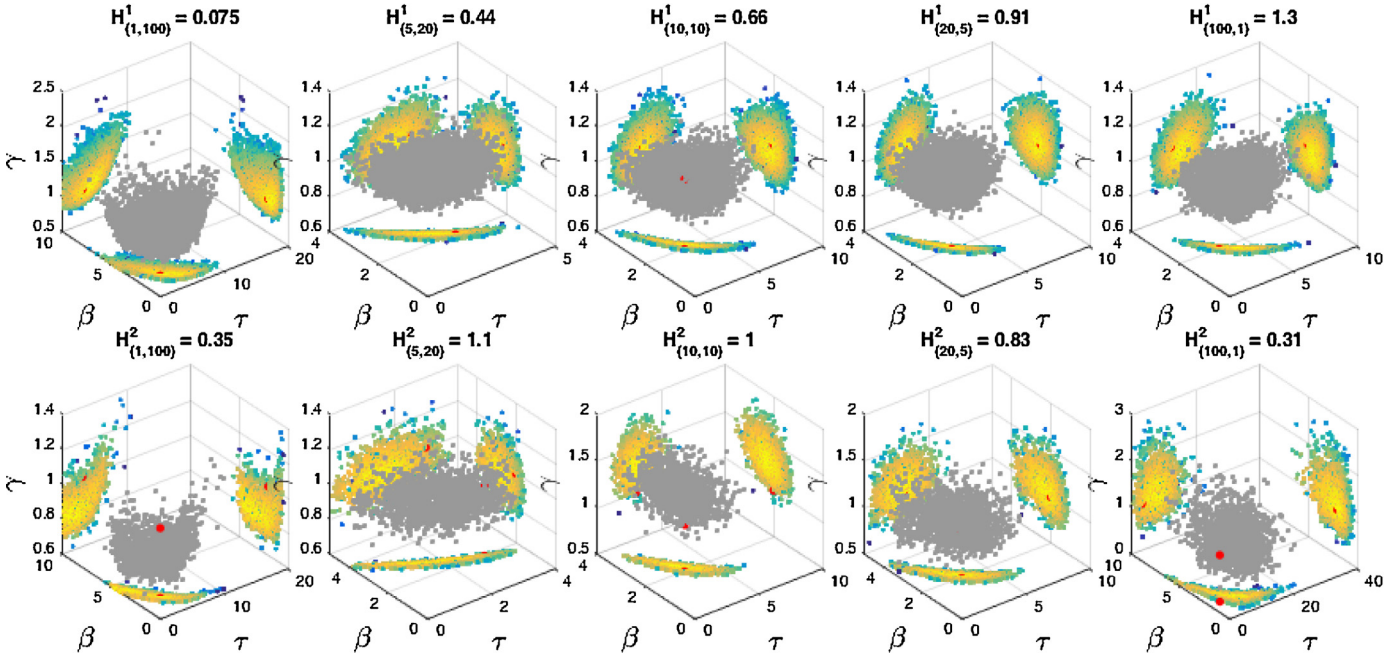

Fig. 3. Posterior densities of the three parameters β, τ and γ for both the cross-sectional (top) and cohort designs (bottom) for the study designs with 100 data points, in 3 dimensions (grey dots) and with bi-dimensional marginals projected in each direction. The information per data point (cross-sectional, $H^1_{(T,K)}$; cohort, $H^2_{(T,K)}$) is reported above each subplot where each 2-tuple, (T,K) represent the number of time points and the number of households sampled, respectively. The red dot represents the true parameters and projected points are coloured based on the likelihood value (blue = low and yellow = high).

lowest values in blue. The top row of each figure shows the results of the experiments for the cross-sectional design while the bottom row shows the cohort design. The design parameters are shown at the top of each subplot i.e. (T, K), referring to the number of time points and households respectively, and the amount of information, (H), as measured by Eq. (14). From each of the figures, we can observe that the likelihood is highest around the true parameter values, shown by the red points, indicating that almost always the true parameters are recovered. Also, there is an inverse correlation between the household transmission rate τ and the random transmission between individuals in the population β as can be seen by the 'banana-like' joint posterior distribution between the two parameters as plotted on the x–y plane in Fig. 3.

These figures show the results using a single dataset generated by a single simulation at baseline parameters i.e. $\boldsymbol{\theta}^*$. In the cross-sectional study, information increases with an increase in the number of time points, with the most information contained in the designs with the highest number of time points. For the cohort design, the most information is contained in $E^2_{(5,20)}$ and $E^2_{(10,100)}$ (see Fig. 3 and S3) for the designs with 100 and 1000 data points respectively. However for the designs with 10 data points (Fig. S2), the best result is observed in the design $E^2_{(1,10)}$ which is the one with the highest number of households.

Better intuition about optimality is drawn from Fig. 4 where we have re-run the analysis with different simulated datasets and recorded the amount of information from each run. Fig. 4 shows the mean and the 95% CI (black solid lines) of the information for all the replicates (dashed lines) and for each of the designs. The number of replicates that we consider are between 10 and 100. Subplots A–C show the information for the designs with 10, 100 and 1000 data points respectively for the cross-sectional design while subplots D, E and F show the same for the cohort design. From this figure, we can observe that designs with more time points contain more information per observation in the cross-sectional design. However, for the cohort study, there exists an intermediate optimum in the study designs giving the most information. As can be seen in subplots E and F in Fig. 4, the optimal designs are (5,20) and (10,100) for the designs with 100 and 1000 data points respectively. As for the designs with 10 data points, there is no evidence to distinguish them as their CIs overlap except for the design (10,1) which contains very little information and therefore it is impossible to distinguish the other four designs meaningfully.

Given that the measure of information about the three parameters is presented in the Shannon information, it is difficult to say how much information we gain for each parameter i.e. which parameters are well estimated depending on the study design. Fig. S4 in the supplementaty material shows the information per datapoint for each of the model parameters β (blue), τ (red) and γ (grey). The left and right hand columns contain the simulations for the cross-sectional and the cohort designs respectively while the rows (from the top) contain the information for the experiments yielding 10, 100 and 1000 data points respectively. In general, the variance in the amount of information for each of the parameters decreases as one increases the number of datapoints from 10 to 1000. Also, τ, which is the within household transmission, seems to be the parameter that is best estimated in almost all of the simulations. Comparing the two bottom subplots, we can also see that we gain more information about the three parameters as we increase the number of time points for the cohort study design (right bottom) compared to the cross sectional design (left bottom).

3.2. Flu-like parameters

Fig. 5 shows the simulations for 100 data points with the Flu-like parameters. As in the previous section, the optimal design for the cross-sectional study is given by the design with the highest frequency of data collection i.e. (100,1). However, the cohort study suggests that the best design is often the one that selects the highest number of households, (1,100). However, we see that for some simulated datasets the presence of an intermediate optimum at (5,20) is restored, highlighting that for complex systems e.g. with multiple transmission levels, non-linear relationships between the

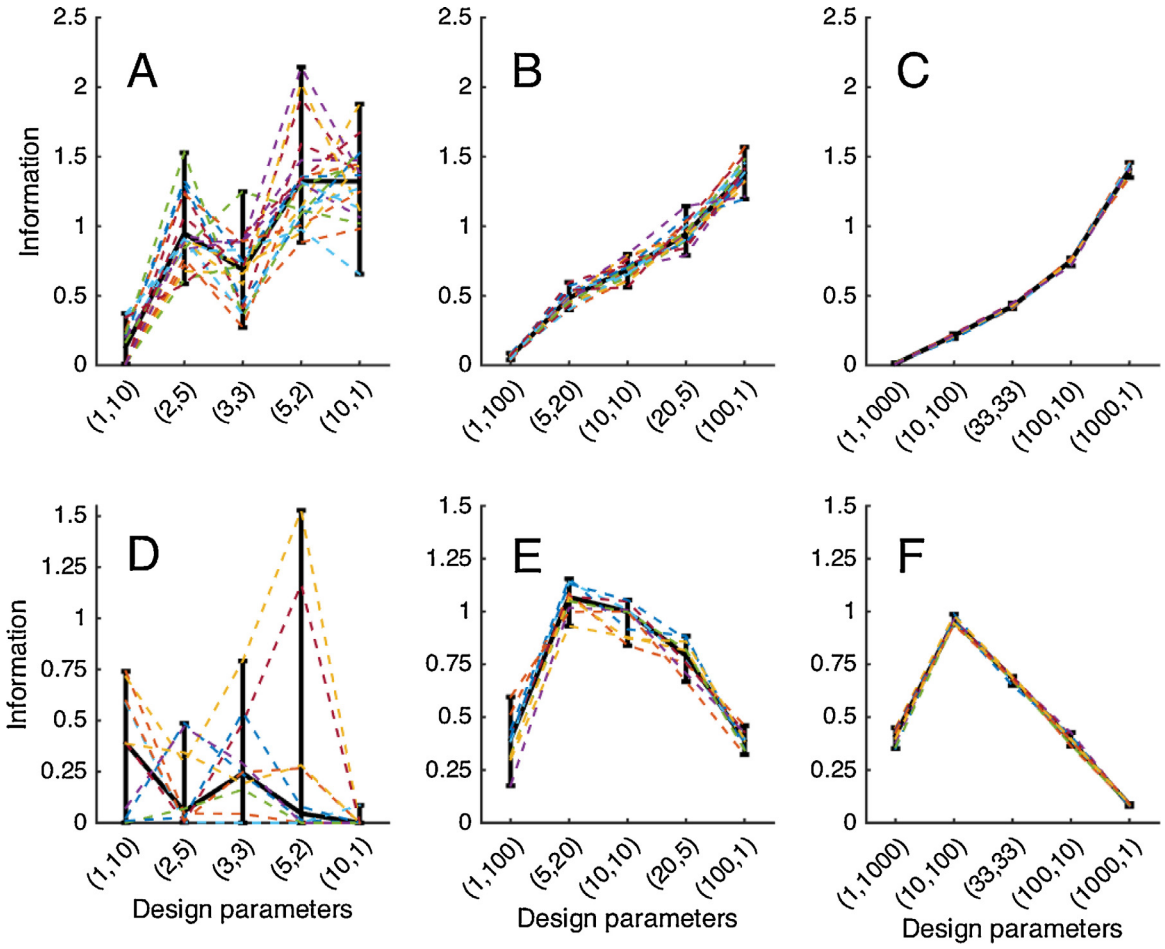

Fig. 4. Comparison of information content for RSV-like parameters. Each of the 2-tuple, (T, K) represent the number of time points and the number of households sampled respectively. (A-C) Cross-sectional and (D-F) cohort designs with (A and D) 10, (B and E) 100 and (C and F) 1000 data points. 15 different simulated datasets are shown as dashed coloured lines, or a black line for the dataset shown in Fig. 3 with the range of values as vertical black bars.

parameters and output interact with the random nature of the simulated data to produce results that are not trivial.

3.3. Systematic sensitivity analysis

We then explored the effect of varying, separately and in combination, both the transmission parameters (β and τ) and the number of individuals in a household (n). Figs. 6 and 7 show the results for the cross-sectional and the cohort studies respectively. The cross-sectional study seems robust to small changes in the parameters values such that the most information per data point is always given by the design with the highest number of time points. However, the cohort study seems to be sensitive to similar changes in the parameter values. For example, subplot D in Fig. 7, which corresponds to the scenarios with the highest community transmission, shows that there exist an intermediate optimal design at (10,10) while all the other scenarios indicate that designs with more households will in general have more information per data point. It is interesting to note that some replicates will have a different optima compared to other replicates within the same set of simulations e.g. Fig. 7B.

4. Discussion

In this work, we have presented a general modelling framework that can be used to make inference on household model parameters based on household-stratified epidemic data. The epidemiological model used is the well studied SIR model and this can be easily modified to reflect the natural progression of any other infection or disease of interest.

The basic idea behind this work is that inference of model parameters can be optimised or improved by selecting different study designs which are used to collect the data. Our results show that, for the cross-sectional study, information increases with an increase in the frequency of sampling i.e. the number of time points at which samples are collected. This is expected as the only within-household information one can collect will be somehow due to the overall epidemic since different households are sampled at each time point. However, for the cohort model, there often exists an intermediate optimum for the designs with 100 and 1000 data points meaning that the best inference of parameters will be the result of a trade-off between the number of households and the frequency of sample collection.

In making a study design decision, the experimenter will need to take other factors about the system being studied into consideration. For example, it might be easier to implement the cohort study as there are fewer households that will need to provide consent for participation compared to the cross-sectional design. Also, if the sampling interval is very short, i.e. intense sampling, there may be limitations as to the timeframe required to obtain consent from a household and enroll it for participation in a study. We remark that it is difficult, given the work presented, to distinguish which of the two protocols (cross-sectional versus cohort) is superior to the other. This is because we assume that all households are the same and therefore any heterogeneities that may arise from

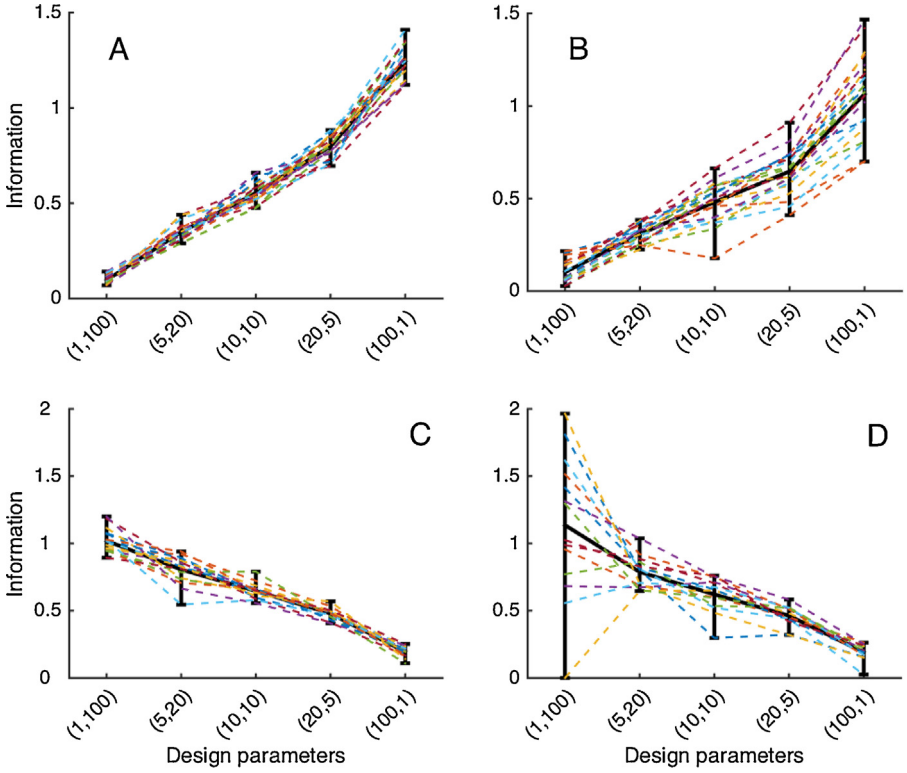

Fig. 5. Comparison of information content for flu-like parameters. Each 2-tuple, (T, K) represent the number of time points and the number of households sampled respectively. (A and B) cross-sectional and (C and D) cohort designs 100 data points. 15 datasets shown as dashed coloured or solid black lines are (A and C) generated from the same parameters (B and D) picked from a normal distribution (centered on the true 'value' and with variance 0.2), with the range of values as vertical black bars.

different households with different characteristics are not captured. It is worth noting that the time points selected for all of the designs always include the final time point and that this is assumed to occur after the epidemic has finished. From the early statistical work of Longini et al. (Longini et al., 1982; Longini and Koopman, 1982), and also more theoretical studies (Ball et al., 1997; Demiris and O'Neill, 2005), we know that information about both the probability of household and community transmission can be estimated from having the final-size distribution of the number of household cases alone. It is therefore expected that the mass of the posterior distributions of β and τ will always concentrate around the baseline values that generated the data for all the designs. Another practical matter worth discussing is the optimal timing of the sampling. For example is it better to rush at the beginning of the epidemic or is it worth waiting and is it even necessary to sample over the entire epidemic. The optimal timing would be dependent on a number of factors among them being the serial interval of the infection which determines on average when a secondary case will

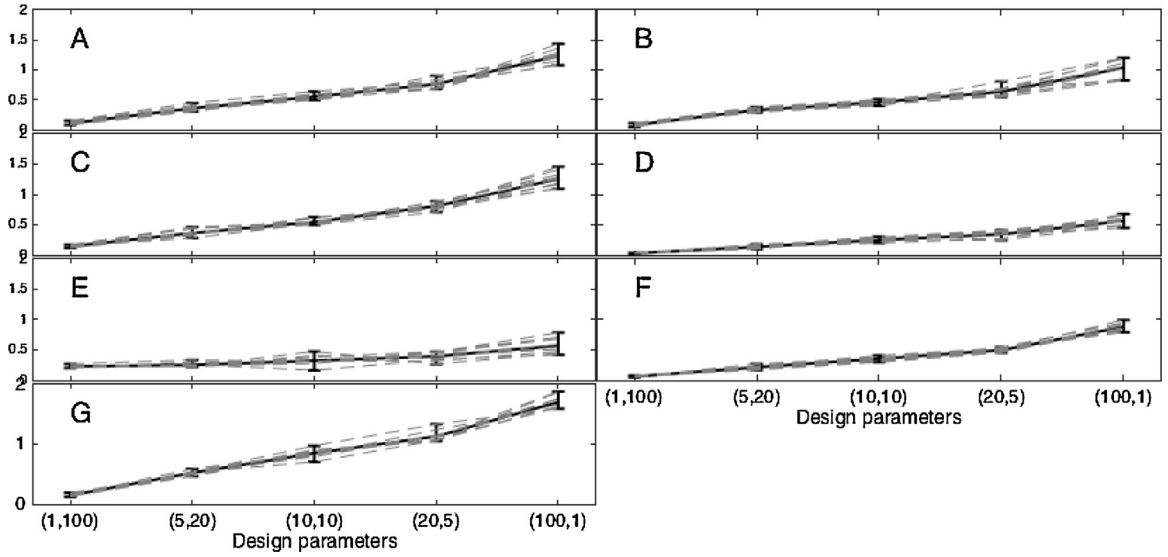

Fig. 6. Comparison of the information per data point generated by different parameter values for the cross sectional design. Each subplot represents a set replicated simulations (dashed lines) per parameter set as shown in Table 1. The solid black lines show the median of the replicates and the vertical bars the region where 95% of the simulations fall.

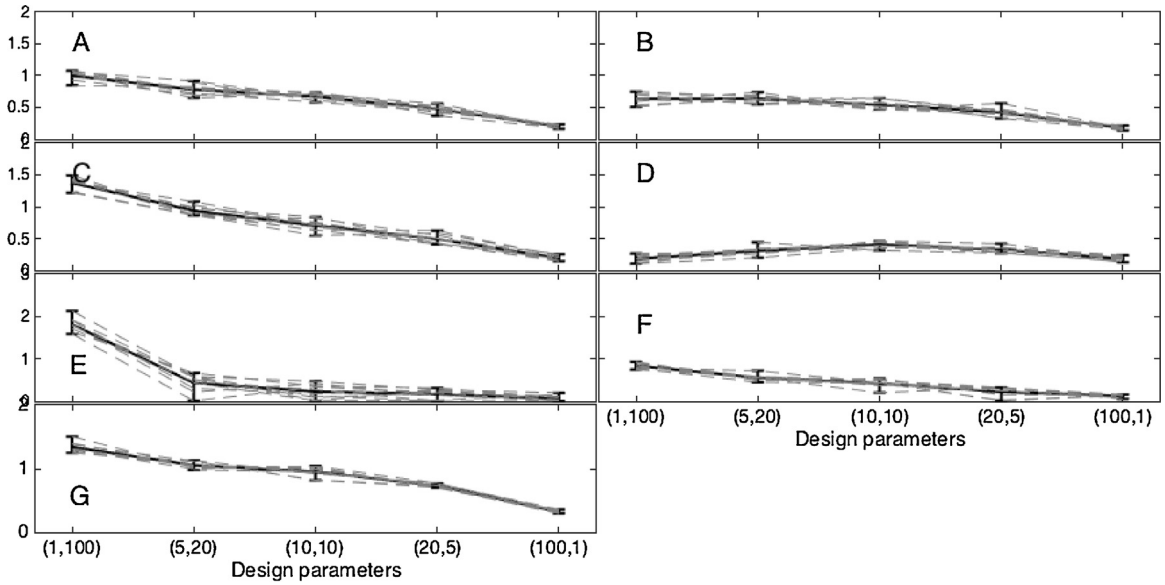

Fig. 7. Comparison of the information per data point generated by different parameter values for the cohort design. Each subplot represents a set replicated simulations (dashed lines) per parameter set as shown in Table 1. The solid black lines show the median of the replicates and the vertical bars the region where 95% of the simulations fall.

start shedding the virus. Also, the probability of virologic confirmation since infection has been shown to be highly dependent on the time since infection at which the sample is taken (Klick et al., 2012; DeVincenzo et al., 2010) consequently influencing the temporal structure of the design. Since we have not explicitly included these two factors in the model, it would be difficult to determine what the optimal timing strategy would be. However, it is clear that a design with more home visits will be less biased than that with less visits and this should come at a cost of greater variance of the parameter estimates due to a reduced sample size for a less intense sampling scheme.

Despite the existence of literature on the optimal design of experiments (Chaloner and Verdinelli, 1995; Cook et al., 2008; Klick et al., 2012, 2014), these methods are not routinely used in the design of studies of infectious disease transmission. This may have been in part due to limitations in computational power. However, with more computational resources available to researchers, we anticipate that these methods will become more commonly applied in the design of field studies in epidemiology. However, certain key questions will still need to be addressed. For example, despite the speed of modern computers, and the fact that our methods would make efficient use of multi-core machines, we were still constrained somewhat by numerical efficiency and future research could fruitfully consider both calculation of the likelihood in a more efficient manner (Ross et al., 2010), as well as improvements to the MCMC scheme (Robert and Casella, 2010). This computational cost has in particular limited the extent of the sensitivity and uncertainty analysis performed. Also, the range of ways in which a study can be designed will need to be taken into considerations. While our simulation-based framework offers a natural way of doing this, it presents a potential challenge in determining an appropriate utility function. The choice of the utility function is usually based on the objective of the experiment. According to Chaloner and Verdinelli (1995), when inference about parameters is the main goal of the study, then Shannon information would be the best measure. However, Shannon information can also be used for prediction and in mixed utility functions that describe multiple simultaneous goals therefore making it quite robust to the objective of the study. It is, or course, possible that the results would change if a different measure was used but that would equally be a reflection of a different study objective. This work has also considered static designs where the experiment is fixed at the beginning of the study. An extension would be to consider the possibility of adaptive designs that change depending on the evolution of the system. This would be a useful feature but probably the most challenging to implement practically given that ethical approval needs to be sought each time the researcher proposes a change to the design.

Despite the challenges above, the kind of studies defined in this work are becoming more common and therefore this work contributes to the discussion of how they should be designed in order to get the most information without collecting unnecessary data that can often be expensive obtain or cause unnecessary risk to participants as some of the specimen collection methods are highly invasive. The fully Bayesian adaptation of our methodology suggested above has utility in such a context as it offers a platform to incorporate what is already known from other experiments in the design process. The experimenter is encouraged to design a different utility function from the one adopted here in order to reflect their study objectives.

Acknowledgments

Work supported by the Engineering and Physical Sciences Research Council. We would like to thank Graham Medley and James Nokes for helpful comments on this manuscript.

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.2016.03.002.

References

Anderson, R.M., May, R.M., 1982. Directly transmitted infectious diseases: control by vaccination. Science 215 (4536), 1053–1060 http://www.jstor.org/stable/1688362.

Anderson, R.M., May, R.M., 1991. Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford/New York.

Baguelin, M., Hoek, A.J.V., 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 (12), 2370–2384, http://dx.doi.org/10.1016/j.vaccine.2010.01.002.
Ball, F., Neal, P., 2002. A general model for stochastic SIR epidemics with two levels of mixing. Math. Biosci. 180, 73–102, http://dx.doi.org/10.1016/S0025-5564(02)00125-6.
Ball, F., Mollison, D., Scalia-Tomba, G., 1997. Epidemics with two levels of mixing. Ann. Appl. Prob. 7 (1), 46–89 http://www.jstor.org/stable/2245132.
Ball, F., 1999. Stochastic and deterministic models for SIS epidemics among a population partitioned into households. Math. Biosci. 156 (1–2), 41–67 http://www.ncbi.nlm.nih.gov/pubmed/10204387.
Becker, N.G., Dietz, K., 1995. The effect of household distribution on transmission and control of highly infectious diseases. Math. Biosci. 127 (2), 207–219.
Becker, N.G., Starczak, D.N., 1997. Optimal vaccination strategies for a community of households. Math. Biosci. 139 (2), 117–132 http://www.sciencedirect.com/science/article/B6VHX-3X2B6TN-B/2/2ac3d7554ee81f84175878e2e1b97296.
Carrat, F., Vergu, E., Ferguson, N.M., Lemaitre, M., Cauchemez, S., Leach, S., Valleron, A.J., 2008. Time lines of infection and disease in human influenza: a review of volunteer challenge studies. Am. J. Epidemiol. 167, 775–785, http://dx.doi.org/10.1093/aje/kwm375.
Cauchemez, S., Ferguson, N.M., Fox, A., Mai, L.Q., Thanh, L.T., Thai, P.Q., Thoang, D.D., Duong, T.N., Minh Hoa, L.N., Tran Hien, N., Horby, P., 2014. Determinants of influenza transmission in South East Asia: insights from a household cohort study in Vietnam. PLoS Pathog. 10 (8), e1004310, http://dx.doi.org/10.1371/journal.ppat.1004310 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4140851&tool=pmcentrez&rendertype=abstract.
Chaloner, K., Verdinelli, I., 1995. Bayesian experimental design: a review. Stat. Sci. 10 (3), 273–304.
Cook, A.R., Gibson, G.J., Gilligan, C.a., 2008. Optimal observation times in experimental epidemic processes. Biometrics 64 (3), 860–868, http://dx.doi.org/10.1111/j.1541-0420.2007.00931.x http://www.ncbi.nlm.nih.gov/pubmed/18047537.
Cowling, B., Chan, K., Fang, V., Chen, K., Fung, O., Wai, W., Sin, J., Seto, W., Yung, R., Chu, W., Chiu, C., Lee, W., Chiu, M., Lee, H., Uyeki, T., Houck, P., Peiris, M., Leung, G., 2009. Facemasks and hand hygiene to prevent influenza transmission in households. A cluster randomized trial. Ann. Intern. Med. 151 (7), 437–446 http://annals.org/article.aspx?articleid=744899&issueno=7.
Demiris, N., O'Neill, P.D., 2005. Bayesian inference for epidemics with two levels of mixing. Scand. J. Stat. 32 (2), 265–280, http://dx.doi.org/10.1111/j.1467-9469.2005.00420.x.
DeVincenzo, J.P., Wilkinson, T., Vaishnaw, A., Cehelsky, J., Meyers, R., Nochur, S., Harrison, L., Meeking, P., Mann, A., Moane, E., Oxford, J., Pareek, R., Moore, R., Walsh, E., Studholme, R., Dorsett, P., Alvarez, R., Lambkin-Williams, R., 2010. Viral load drives disease in humans experimentally infected with respiratory syncytial virus. Am. J. Respir. Crit. Care Med. 182 (10), 1305–1314, http://dx.doi.org/10.1164/rccm.201002-0221OC http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3001267&tool=pmcentrez&rendertype=abstract.
Frank, A.L., Taber, L.H., Wells, C.R., Wells, J.M., Glezen, W.P., P.A., 1981. Patterns of shedding of myxoviruses and paramyxoviruses in children. J. Infect. Dis. 144 (5), 433–441.
Glezen, W.P., Taber, L.H., Frank, A.L., Kasel, J.A., 1986. Risk of primary infection and reinfection with respiratory syncytial virus. Am. J. Dis. Child. 140, 543–546.
Hadeler, K.P., Castillo-Chavez, C., 1995. A core group model for disease transmission. Math. Biosci. 128 (1–2), 41–55 http://www.sciencedirect.com/science/article/B6VHX-3XY2KCX-3/1/3fdd0120002d47d0eff866a2bde40bd9.
Hayward, A.C., Fragaszy, E.B., Bermingham, A., Wang, L., Copas, A., Edmunds, W.J., Ferguson, N., Goonetilleke, N., Harvey, G., Kovar, J., Lim, M.S.C., McMichael, A., Millett, E.R.C., Nguyen-Van-Tam, J.S., Nazareth, I., Pebody, R., Tabassum, F., Watson, J.M., Wurie, F.B., Johnson, A.M., Zambon, M., 2014. Comparative community burden and severity of seasonal and pandemic influenza: results of the Flu Watch cohort study. Lancet Respir. Med. 2 (6), 445–454, http://dx.doi.org/10.1016/S2213-2600(14)70034-7 http://www.ncbi.nlm.nih.gov/pubmed/24717637.
Horby, P., Mai, L.Q., Fox, A., Thai, P.Q., Thi Thu Yen, N., Thanh, L.T., Le Khanh Hang, N., Duong, T.N., Thoang, D.D., Farrar, J., Wolbers, M., Hien, N.T., 2012. The epidemiology of interpandemic and pandemic influenza in Vietnam, 2007–2010: the Ha Nam household cohort study I. Am. J. Epidemiol. 175 (10), 1062–1074, http://dx.doi.org/10.1093/aje/kws121 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3353138&tool=pmcentrez&rendertype=abstract.
House, T., Keeling, M.J., 2008. Deterministic epidemic models with explicit household structure. Math. Biosci. 213 (1), 29–39.
House, T., Keeling, M.J., 2009. Household structure and infectious disease transmission. Epidemiol. Infect. 137 (5), 654–661.
House, T., Inglis, N., Ross, J.V., Wilson, F., Suleman, S., Edeghere, O., Smith, G., Olowokure, B., Keeling, M.J., 2012. Estimation of outbreak severity and transmissibility: Influenza A(H1N1)pdm09 in households. BMC Med. 10 (1), 117, http://dx.doi.org/10.1186/1741-7015-10-117 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3520767&tool=pmcentrez&rendertype=abstract.
Keeling, M.J., Rohani, P., 2007. Modeling Infectious Diseases in Humans and Animals. Princeton University Press.
Klick, B., Leung, G., Cowling, B., 2012. Optimal design of studies of influenza transmission in households. I: Case-ascertained studies. Epidemiol. Infect. 140 (1), 106–114, http://dx.doi.org/10.1017/S0950268811000392.Optimal http://journals.cambridge.org/abstract_S0950268811000392.
Klick, B., Nishiura, H., Leung, G.M., Cowling, B.J., 2014. Optimal design of studies of influenza transmission in households II: comparison between cohort and case-ascertained studies. Epidemiol. Infect. 142 (4), 744–752, http://dx.doi.org/10.1016/j.biotechadv.2011.08.021.Secreted, arXiv:NIHMS150003.
Lindley, D.V., 1970. Bayesian Statistics, A Review, Society for Industrial and Applied Mathematics, Bristol.
Longini Jr., I.M., Koopman, J., 1982. Household and community transmission parameters from final distributions of infections in households. Biometrics 38 (1), 115–126 http://www.jstor.org/stable/2530294.
Longini Jr., I.M., Koopman, J.S., Monto, A.S., Fox, J.P., 1982. Estimating household and community transmission parameters for influenza. Am. J. Epidemiol. 115 (5), 736–751.
Mathworks Inc., 2014. MATLAB R2014b, Version 7. 7. 0.
Munywoki, P., Koech, D., 2013. The source of respiratory syncytial virus infection in infants: a household cohort study in rural Kenya. J. Infect. Dis., http://dx.doi.org/10.1093/infdis/jit828.
Neal, P., 2012. Efficient likelihood-free Bayesian computation for household epidemics. Stat. Comput. 22 (6), 1239–1256.
Okiro, E.A., White, L.J., Ngama, M., Cane, P.A., Medley, G.F., Nokes, D.J., 2010. Duration of shedding of respiratory syncytial virus in a community study of Kenyan children. BMC Infect. Dis. 10, 15.
O'Neill, P.D., Balding, D.J., Becker, N.G., Eerola, M., Mollison, D., 2000. Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. J. R. Stat. Soc.: Ser. C (Appl. Stat.) 49 (4), 517–542.
Poletti, P., Merler, S., Ajelli, M., Manfredi, P., Munywoki, P.K., Nokes, J.D., Melegaro, A., 2015. Evaluating vaccination strategies for reducing infant respiratory syncytial virus infection in low-income settings. BMC Med. 13 (1), 1–11, http://dx.doi.org/10.1186/s12916-015-0283-x http://www.biomedcentral.com/1741-7015/13/49.
Read, J.M., Eames, K.T., Edmunds, W.J., 2008. Dynamic social networks and the implications for the spread of infectious disease. J. R. Soc. Interface 5 (26), 1001–1007.
Robert, C.P., Casella, G., 2010. Introducing Monte Carlo Methods with R. Springer.
Ross, J.V., House, T., Keeling, M.J., 2010. Calculation of disease dynamics in a population of households. PLoS ONE 5 (3), e9666, http://dx.doi.org/10.1371/journal.pone.0009666.
Schenzle, D., 1984. An age-structured model of pre- and post-vaccination measles transmission. Math. Med. Biol. 1 (2), 169–191 http://imammb.oxfordjournals.org/cgi/content/abstract/1/2/169.
Shannon, C., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 3 (27), 379–423.
Sutton, A.J., House, T., Hope, V.D., Ncube, F., Wiessing, L., Kretzschmar, M., 2012. Modelling HIV in the injecting drug user population and the male homosexual population in a developed country context. Epidemics 4 (1), 48–56, http://dx.doi.org/10.1016/j.epidem.2011.12.001 http://www.ncbi.nlm.nih.gov/pubmed/22325014.