Bayesian inference for multi-strain epidemics with application to Escherichia coli O157

Touloupou, Panayiota; Finkenstadt, Barbel; Besser, Thomas E. ; French, Nigel P.; Spencer, Simon E. F.

DOI:
10.1214/20-AOAS1366

Submitted to the Annals of Applied Statistics

BAYESIAN INFERENCE FOR MULTI-STRAIN EPIDEMICS WITH APPLICATION TO *ESCHERICHIA COLI* O157:H7 IN FEEDLOT CATTLE

By Panayiota Touloupou[*,‡], Bärbel Finkenstädt[‡], Thomas E. Besser[§], Nigel P. French[¶] and Simon E.F. Spencer[†,‡]

University of Warwick [‡], Washington State University[§] and Massey University[¶]

For most pathogens testing procedures can be used to distinguish between different strains with which individuals are infected. Due to the growing availability of such data, multi-strain models have increased in popularity over the past few years. Quantifying the interactions between different strains of a pathogen is crucial in order to obtain a more complete understanding of the transmission process, but statistical methods for this type of problem are still in the early stages of development. Motivated by this demand, we construct a stochastic epidemic model, that incorporates additional strain information, and propose a statistical algorithm for efficient inference. The model improves upon existing methods in the sense that it allows for both imperfect diagnostic test sensitivities and strain misclassification. Extensive simulation studies were conducted in order to assess the performance of our method, while the utility of the developed methodology is demonstrated on data obtained from a longitudinal study of *Escherichia coli* O157:H7 strains in feedlot cattle.

1. Introduction. Over the past decades, mathematical models have been established as an important tool for understanding the transmission dynamics of infectious diseases. Statistical inference in transmission dynamic models is not trivial and requires specialised methodology. A key facet of the problem is that the actual process of infection is in most cases only partially observed, in the sense that the times of acquiring and clearing infection are not directly observed. This is because individuals are often tested at sparse time points and the diagnostic tests that are used to detect the disease are typically imperfect. A further complication is the existence of dependencies in the data that arises because of the contacts made between individuals. For all these reasons, it is often difficult to analytically evaluate the likelihood function on which inferences are based, since its calculation involves integrating out all unobserved quantities.

When the full data are available, i.e. the times of infection and recovery are known, one can use standard techniques to obtain estimates of the parameters of interest, for example by maximum likelihood methods (Becker, 1989). However, since data are typically incomplete, most of the literature deals with methods that tackle the problem of inference in partially observed epidemics. Many exact and approximate approaches have been developed; for an extended review see for example Becker (1989), Daley and Gani

[*]Supported by a University of Warwick Department of Statistics PhD scholarship

[†]Supported by MRC grant MR/P026400/1 and EPSRC grant EP/R018561/1

Keywords and phrases: Multi-state Markov model, Misclassification, Epidemiology, Markov chain Monte Carlo, Genotypes

(2001), O'Neill (2002) and Diekmann, Heesterbeek and Britton (2012), among others. Initial approaches use martingale theory to obtain method of moments estimates for the model parameters (Becker, 1989; Rida, 1991; Becker and Hasofer, 1997), but it is hard to extend these methods to the complex models that are used in practice. Instead, it is more common to employ data augmentation methods, in which the missing data are treated as additional model parameters. For example, Becker (1997) and Becker and Britton (1999) tackle the problem with an Expectation-Maximisation (EM) algorithm. An alternative approach is the use of Markov chain Monte Carlo (MCMC) methods under the Bayesian paradigm, which are currently popular techniques for analysing data on partially observed infectious diseases.

The first data augmentation MCMC algorithms were developed by Gibson and Renshaw (1998) and O'Neill and Roberts (1999) for statistical analysis of the continuous time SEIR and SIR models, respectively. After that, several works adapting MCMC techniques with data imputation appeared in the literature, including Auranen et al. (2000), O'Neill (2002), Morton and Finkenstädt (2005), Jewell et al. (2009), Kypraios et al. (2010), Erästö, Hoti and Auranen (2012), Spencer et al. (2015) and numerous other papers. O'Neill and Becker (2001) and Streftaris and Gibson (2004) were among the first to apply MCMC in models with a non-Markovian infection period. Smith and Vounatsou (2003) demonstrate the use of discrete time hidden Markov models for modelling longitudinal epidemiological data. The framework extends to partially observed continuous time epidemic models, see for example Fearnhead and Meligkotsidou (2004). Some of the literature focuses on coupled hidden Markov models for modelling partially observed longitudinal household data, which can effectively account for interactions between individuals and are used in the present work (for example, Dong, Pentland and Heller, 2012; Touloupou, Finkenstädt and Spencer, 2019).

Another class of techniques that have growing popularity in several scientific fields, along with epidemiology, are the so-called simulation-based methods. These include Approximate Bayesian Computation (ABC; McKinley, Cook and Deardon, 2009; Neal, 2012; Kypraios, Neal and Prangle, 2017), iterated filtering for maximum likelihood estimation in partially observed epidemic models (Ionides, Bretó and King, 2006; Ionides et al., 2015), Sequential Monte Carlo ABC (Toni et al., 2009) and pseudo-marginal methods (McKinley et al., 2014). Clancy and O'Neill (2007) demonstrated the usefulness of rejection sampling as an alternative to MCMC. Lastly, there are examples of studies exploring the use of non-parametric methods for inference in epidemic models (Xu, Kypraios and O'Neill, 2016; Kypraios et al., 2018).

Our work is concerned with epidemic data containing information regarding the strain of a pathogen with which individuals are colonised. When multiple strains of a pathogen are coexisting then the number of infectious states an individual can exhibit is greatly increased and existing inference approaches become prohibitive. Here we utilise recent advances in computationally scalable data augmentation (Touloupou, Finkenstädt and Spencer, 2019) to impute the missing strain data and hence to perform inference for a pathogen with multiple strains.

In this study a strain is defined as "a genetically distinct and identifiable subpopulation of a parasite or pathogen which has distinct epidemiological, immunological or patholog-

ical characteristics" (Lord et al., 1999). As such we are interested in modelling relatively low resolution genotype data, such as Pulsed Field Gel Electrophoresis (PFGE) or multi-locus sequence typing data, or high resolution data (e.g. whole genome sequencing data) on which some genetic clustering has been performed. We consider each strain as a distinct pathogen population circulating within a population of hosts and ignore the effects of evolutionary processes acting during the study. In such cases, it is reasonable to examine whether there is appreciable heterogeneity between the different strains in, for example, their transmissibility or the duration for which each strain remains within the host. Additionally, we would like to address the question of between-strain competition, that is if carriage of a certain strain reduces the possibility of being colonised by a different strain. Such knowledge can further our understanding of the epidemiology of an infectious disease.

Parameter estimation in a multi-strain pathogen context may be challenging due to identifiability issues, which occur due to the many strain-specific parameters that need to be estimated. A pragmatic solution can be to group multiple strains into a single class in order to simplify the model. Some examples include Cauchemez et al. (2006) who group strains according to whether they are included in the vaccine formulation or not, Erästö, Hoti and Auranen (2012) who classify strains according to their frequency in the data and Melegaro et al. (2007) who use a separate model for each strain. More recently, Worby et al. (2016) use genome sequence information to classify the isolates into genetically similar groups. In all of these approaches, the problem of missing data is dealt with either by extending the Bayesian data augmentation framework proposed by Gibson and Renshaw (1998) and O'Neill and Roberts (1999) (Cauchemez et al., 2006; Erästö, Hoti and Auranen, 2012; Worby et al., 2016), or by adopting a maximum profile likelihood approach (Melegaro et al., 2007). An additional complication arises from the fact that strain information often relies on diagnostic tests which suffer from low sensitivity. As a result, carriage incidents may remain undetected or be recorded as the wrong strain. Most of the above methods assume that test results are observed without error and hence do not allow the possibility of false positive or false negative outcomes. The exception is Worby et al. (2016) who estimate a common test sensitivity for all groups. However, their model does not allow for competition between strains and is therefore unable to separate strain misclassification from strain replacement. Our model addresses some of the limitations of the existing approaches, by simultaneously allowing for imperfect test sensitivities and strain misclassification.

The paper is structured as follows. We start by describing our motivating dataset obtained from a longitudinal study of *Escherichia coli* O157:H7 in cattle. *E. coli* O157:H7 is a human pathogen that causes severe disease symptoms in people but does not cause disease in its cattle reservoir host. The novel dataset is presented in Section 2.1, in which the diagnostic tests not only show whether an individual carries the disease, but also provide additional information regarding the genotypes in which the bacterium appears. The transmission and observation model are formulated in Sections 2.2 and 2.3, respectively. In Section 2.4 we describe the algorithm which is used for posterior inference. Performance of our method is assessed on simulated data under different scenarios in Section 3.1. In Section 3.2 we apply the proposed methodology to the dataset described in Section 2.1, in order to further our understanding regarding the dynamics of various genotypes of

E. coli O157:H7 in cattle, as well as to investigate between-genotype competition. Finally, in Section 4 we conclude with a discussion.

2. Material and methods.

2.1. *Colonisation and pathogen genetic data.* A longitudinal study of natural rectoanal junction colonisation and faecal excretion of *E. coli* O157:H7 was conducted in feedlot cattle. In this study 160 cattle were randomly assigned to twenty pens; eight animals each. The pens were separated by an empty pen, ensuring that no direct contact was possible between animals of different pens. In addition, each pen had an individual water supply and a separate feed bunk. The animals were housed in North and South pens measuring 6m×17m and 6m×37m, respectively.

Animals were sampled approximately twice per week over a 14-week period. In brief, at each sampling date two samples were collected from each animal: a recto-anal mucosal swab (RAMS) sample and a sample of freshly passed manure. In the original study the samples were cultured in selective media and a polymerase chain reaction (PCR) test was used to identify whether or not the samples contained *E. coli* O157:H7. The cultures were then frozen for future use. This dataset was originally described by Cobbold et al. (2007) and modelled by Spencer et al. (2015) who used a Susceptible-Infected-Susceptible (SIS) transmission model for the spread of infection in a pen, in which each individual is assumed to belong to one of two states: either susceptible or colonized (infected).

The cultures were subsequently unfrozen and PFGE was used to identify the genotypes within a subset of the *E. coli* O157:H7 cultures. In this paper we introduce the additional genotyping data that has not previously been published or analysed. More specifically, a subset of 12 PCR positive samples (either RAMS or faecal) were randomly selected from each pen for genotyping using PFGE as described by Tenover et al. (1997). For 5 pens fewer genotyped samples were available, either because fewer than 12 positive samples were obtained or genotyping was unsuccessful. Overall, there were a total of 223 genotyped samples among the 756 positive samples, a proportion of almost 30%.

A total of 48 different genotypes were identified in the study population which we assign arbitrary labels according to the order in which they appeared in the PFGE dendrogram. Seven genotypes were observed in at least 10 RAMS and/or faecal samples, whilst the remaining 41 genotypes were detected in at most 5 isolates with 24 appearing only once. Figure 1 illustrates the frequencies of genotypes, ranked according to the order in which they appeared in the dendrogram. In addition, Table 1 summarises the frequencies of the 7 most common *E. coli* O157:H7 PFGE genotypes (D, J, X, b, d, f and l) in the data by pen groups, i.e. North and South.

Among the 160 cattle examined in the study, 106 (66.25%) gave at least one genotyped sample. For these, the median number of genotyped samples was 2 (min-max: 1-9). Figure 2 presents data collected in a subset of 4 pens. These data allow us to comment on the micro-epidemics of a genotype within a pen. For example, at the beginning of the study genotype X was detected in the samples collected from individual 5 in pen number 8, and then a micro-epidemic was observed with at least 5 individuals carrying genotype X during the following period. Note that, several individuals were never selected for genotyping (e.g. animal 7, pen 8).

FIG 1. *Frequencies of genotypes identified in E. coli O157:H7 data, ordered by their appearance in the PFGE dendogram. Strains* $1, 2, \ldots, 7$ *are defined for the 7 most common genotypes recovered in this dataset, types D, J, X, b, d, f and l respectively, and the remaining genotypes are treated as a single strain, referred as type 8.*

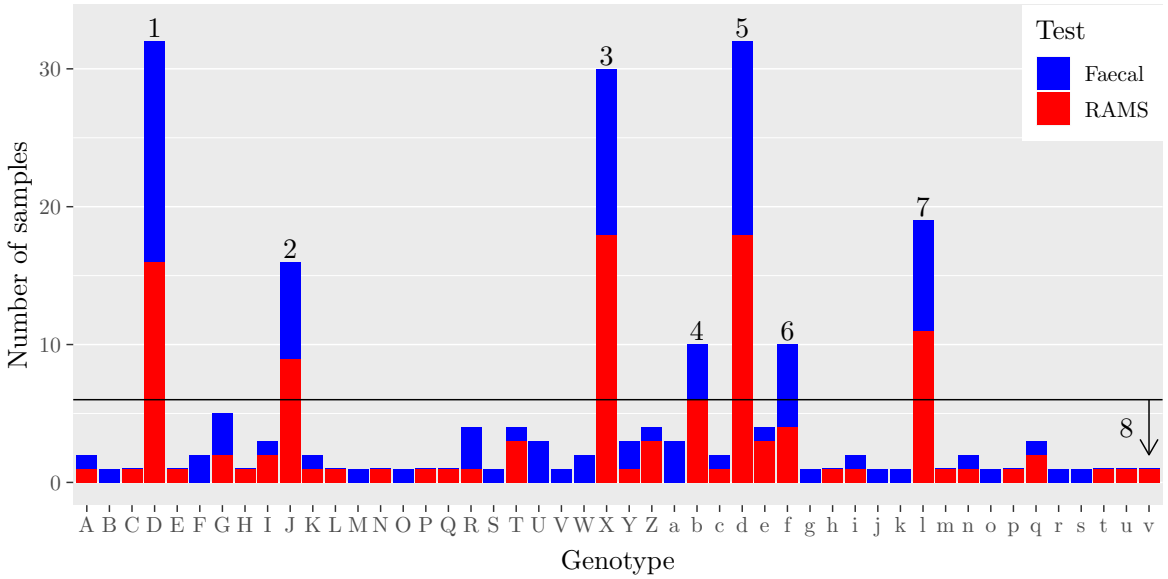


TABLE 1
Distribution of E. coli O157:H7 observed genotypes during follow-up of 160 cattle among 12 North and 8 South pens.

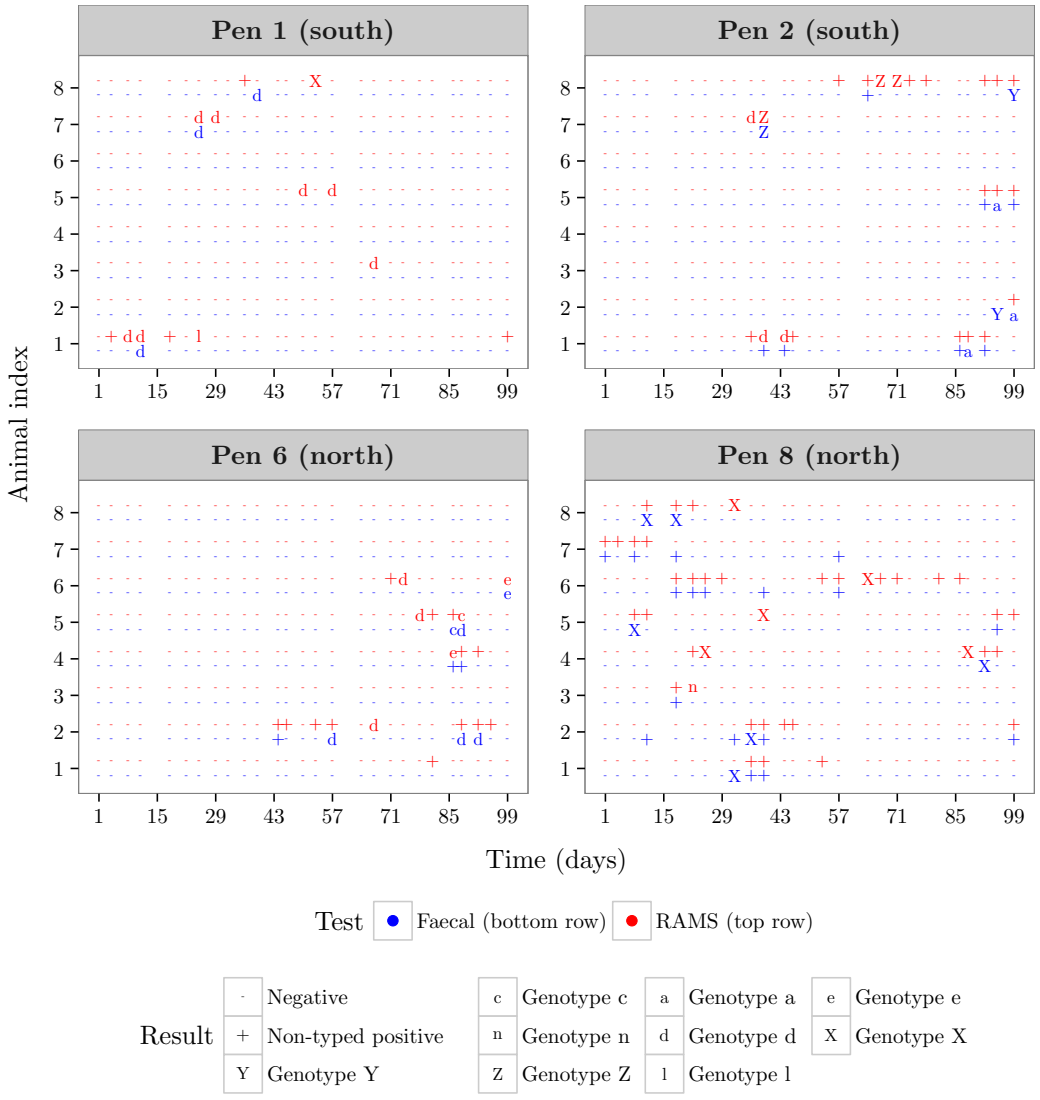
Genotypes	**Observed positive samples**					
	South		**North**		**Total**	
	No.	**%**	**No.**	**%**	**No.**	**%**
D (1)	10	3.6	22	4.6	32	4.2
J (2)	2	0.7	14	2.9	16	2.1
X (3)	2	0.7	28	5.8	30	4.0
b (4)	3	1.1	7	1.5	10	1.3
d (5)	15	5.5	17	3.5	32	4.2
f (6)	3	1.1	7	1.5	10	1.3
l (7)	19	6.9	0	0.0	19	2.5
Other genotypes (8)	32	11.6	42	8.7	74	9.8
Non-typed[†]	189	68.7	344	71.5	533	70.5
Total	275	100	481	100	756	100

[†] Positive samples that were not chosen to be genotyped.

Moreover, of the 223 genotyped samples, 22 pairs of positive samples were chosen to be genotyped, where a pair is defined as RAMS and faecal isolates from the same individual on the same sampling date. Of these, there were 19 occasions in which an animal was observed to carry the same genotype by RAMS and faecal isolates, and the remaining 4

were pairs of different genotypes (e.g. animal 5, pen 6 at day 88); this could be attributed to misclassification errors of the genotyping procedure, or it could be evidence of co-infection.

FIG 2. *RAMS and faecal samples for each individual (top red and bottom blue respectively) collected in pens 1, 2, 6 and 8 participating in the study. "–" indicates negative sample, "+" indicates that the sample was positive but not chosen for genotyping; otherwise, the genotype name is given.*



2.2. *Transmission model.* The unobserved (hidden) colonisation process within each pen is modelled as a multi-state discrete time Markov model. This model is an extension of the standard individual-based SIS model (Anderson and May, 1991), but also incorporates strain-specific information. More precisely, we define a discrete time Markov transition

model with n_g+1 states, in which individuals belong to a state according to their carriage status. The possible states include being a non-carrier (state 0) or being a carrier of one of the n_g strains (states $1, 2, \ldots, n_g$). The model assumes that an individual can carry at most one strain at a time: when individuals acquire a new strain then it replaces the existing strain. We do not model co-infection, which can be justified by the fact that there were only four occasions in the data set in which an individual was observed to carry different genotypes on RAMS and faecal positive samples taken on the same sampling day. These contradictory pairs can be handled by allowing for the possibility of genotype misspecification, as described in the subsequent Section 2.3.

The transition probabilities are defined in Equation (2.3) based on the following functions, which we call 'transition rates' because of their similarity to the transition rates in the analogous continuous time Markov process. The transition rate between any two carriage states in this Markov model, $r, s \in \{0, 1, \ldots, n_g\}$, for each individual in pen p at day t, is defined for three cases:

$$h^p_{r,s}(t) = \begin{cases} \lambda^p_s(t) & r = 0, s \neq 0; \text{ colonisation,} \\ \delta\, \lambda^p_s(t) & r, s > 0 \text{ and } r \neq s; \text{ change of genotype,} \\ \mu_r & r \neq 0, s = 0; \text{ clearance,} \end{cases} \tag{2.1}$$

where the first case defines the colonisation rate at which a non-carrier acquires a particular genotype s ($0 \mapsto s$) at day t, for which the rate depends on the genotype, the day and the individual's pen. The second case corresponds to the rate of transition from carriage of genotype r to carriage of genotype s ($r \mapsto s$), where $r \neq s$. Between-genotype competition in colonising the host is included in the model by using an additional parameter $\delta > 0$ to scale the rate of colonisation in an individual already carrying another genotype. This parameter is assumed to be the same for all genotype pairs. Finally, once colonised, individuals can recover from carriage of genotype r ($r \mapsto 0$) according to a genotype-dependent clearance rate μ_r that is constant over time and across different pens. A simplified version of this model is presented in Figure 3, with only three genotypes. Since individuals were assigned to pens at random, we assume that at the beginning of the study each individual is colonised by genotype s independently with a genotype-dependent probability ν_s.

In addition, the model assumes that the rate at which a non-carrier individual acquires a genotype is pen-, type- and time-dependent, varying as a function of the number of other pen members carrying this particular genotype. To be more specific, for a non-carrying individual in pen p, where $p \in \mathcal{N}$ (North group) or $p \in \mathcal{S}$ (South group), the rate of colonisation of genotype s, at any given time t, is defined as the sum of two components as follows:

$$\lambda^p_s(t) = \alpha_s + \left(\mathbb{1}_{\{p \in \mathcal{S}\}} + \gamma\, \mathbb{1}_{\{p \in \mathcal{N}\}}\right) \beta_s\, I^p_s(t-1), \tag{2.2}$$

where $I^p_s(t)$ denotes the number of carriers of genotype s in pen p at time t and $\mathbb{1}$ denotes the indicator function. The genotype-specific terms β_s and α_s represent the rates of colonisation from contacts with other members of the pen (within-pen colonisation rate) and from sources outside of the pen (external colonisation rate), respectively. To account for

differences between North (smaller) and South (larger) pens the within-North pen colonisation rates are multiplied with γ, where γ is the relative acquisition rate in smaller versus bigger pens, as shown in Equation (2.2). This can be justified by differences in pen sizes (6m×17m compared with 6m×37m) and by previous finding in Spencer et al. (2015) and Touloupou (2016) that animals in smaller pens are at greater risk of within-pen infection.

The proportion of positive samples in the study population was less than 10%, with only a few events per genotype (Figure 1). Consequently, analysing these data using a multi-state model where the possible states include being a carrier of one of the 48 different genotypes, presents a considerable challenge; the large number of genotype-specific parameters lead to problems in identifiability. Like most of the previous epidemic analyses, we solve this problem by dividing the genotypes into groups as follows. States of carriage are defined for the 7 genotypes most commonly recovered in this study, types D=1, J=2, X=3, b=4, d=5, f=6 and l=7. The remaining genotypes are treated as a single group, referred as the "Pooled" group, and assumed to be of the same type 8. This unrealistic assumption affects a small proportion (9.8%) of the genotyped samples. For the most common genotypes we assume their own individual rates of acquisition and clearance, and the pooled group has its own rate parameters. The colonisation process parameters are described in Table 2.

Formulating the model with a pooled group substantially reduces the number of different carriage states to 9, with $n_g = 8$. To this end, we denote the carriage state of individual $c \in \{1, 2, \ldots, C\}$ in pen $p \in \{1, 2, \ldots, P\}$ on day $t \in \mathcal{T}^{c,p}$, by $X_t^{[c,p]} \in \mathcal{X}_g = \{0, 1, \ldots, n_g\}$, where $X_t^{[c,p]} = 0$ refers to the non-carriage state, state $X_t^{[c,p]} = n_g$ to carriage of the pooled group, and state $X_t^{[c,p]} = s$, for $0 < s < n_g$, to carriage of one of the common genotypes.

FIG 3. *The model graph for an individual that belongs to pen p in which, for simplicity, three genotypes are considered, denoted as 1, 2 and 3 respectively, and four carriage states. Transitions between the states are governed by rates of acquisition and clearance, as marked at each arrow. The acquisition rates depend on the number of individuals within the pen carrying that particular genotype, and for individuals already carrying another genotype the rates are adjusted by a competition parameter δ. Moreover, the rates of within-pen acquisition for individuals that belong to a smaller North pen are scaled by a factor γ.*

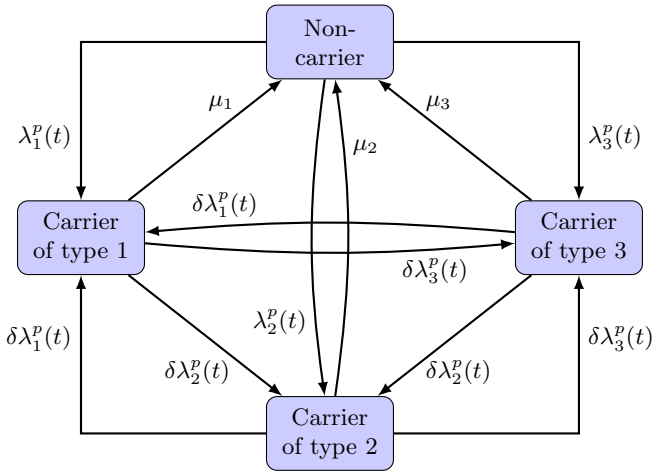


TABLE 2
Symbols and interpretations of the colonisation process parameters, where $s = 1, 2, \ldots, 8$ denotes the genotype.

Parameter	Interpretation
α_s	External colonisation rate for genotype s (days^{-1})
β_s	Within-pen colonisation rate for genotype s (days^{-1})
μ_s	Clearance rate for genotype s (days^{-1})
ν_s	Initial probability of carriage with genotype s
δ	Relative colonisation rate in a carrier versus non-carrier individual
γ	Relative colonisation rate in smaller versus bigger pens

The observation (sampling) period for each individual is defined as the period from the first sample to the last one, denoted by $\mathcal{T}^{c,p} \subseteq \{1, 2, \ldots, T\}$, where the first sample is taken at $t = 1$ and the last at $t = T$.

According to the assumptions and notation above, the model is defined as a discrete-time Markov process with time interval equal to one day (the greatest common divisor of the times between sampling events), in which the current status of each individual depends on the previous status of all the individuals within the pen. The probabilities of transition between states, for any individual c in pen p at time t, can be arranged in a $(n_g + 1) \times (n_g + 1)$ matrix $\mathbf{M}^p(t)$ (time- and pen-dependent) with elements $m^p_{r,s}(t)$, for $r, s = 0, 1, 2, \ldots, n_g$ and $t \in \mathcal{T}^{c,p} \setminus \{1\}$. For convenience we start indexing the rows and columns of $\mathbf{M}^p(t)$ from 0. The off-diagonal elements of $\mathbf{M}^p(t)$ are specified below,

$$
\begin{aligned}
m^p_{r,s}(t) &= \mathbb{P}\left(X_t^{[c,p]} = s \middle| X_{t-1}^{[c,p]} = r, \mathbf{X}_{t-1}^{[-c,p]}\right) \\
&= \underbrace{\left(1 - \exp\left(-\sum_{\substack{j=0 \\ j\neq r}}^{n_g} h^p_{r,j}(t)\right)\right)}_{\text{Probability that an event occurs}} \times \underbrace{\frac{h^p_{r,s}(t)}{\sum_{\substack{j=0 \\ j\neq r}}^{n_g} h^p_{r,j}(t)}}_{\substack{\text{Probability that} \\ \text{event } r \mapsto s \text{ occurs,} \\ \text{given an event occurs}}},
\end{aligned} \tag{2.3}
$$

for $r \neq s$, where $\mathbf{X}_{t-1}^{[-c,p]}$ is the vector of the hidden states of the remaining individuals within pen p at time $t-1$. Diagonal elements in $\mathbf{M}^p(t)$ contain the $m^p_{r,r}(t)$, which are defined as $m^p_{r,r}(t) = 1 - \sum_{\substack{j=0 \\ j\neq r}}^{n_g} m^p_{r,j}(t)$ so that the sum of all elements in each row equals one. Thus, using this parametrization the transition probability in the case where $r = s$ is given by $m^p_{r,r}(t) = \exp\left(-\sum_{\substack{j=0 \\ j\neq r}}^{n_g} h^p_{r,j}(t)\right)$, which is equal to the probability of there being no events in a Poisson process with rate $\sum_{\substack{j=0 \\ j\neq r}}^{n_g} h^p_{r,j}(t)$.

2.3. *Observation model.* In our study, the diagnostic tests used to detect *E. coli* O157:H7 in cattle are imperfect; the sensitivities of these techniques may be as low as 50% and thus

some colonised individuals remain undetected (see for example Spencer et al., 2015). In addition, the PFGE clustering of the data used for genotyping the samples was assumed to have less than perfect accuracy, meaning that the carriage states may have not been recorded with their true genotype. Therefore the classification at an observation time can sometimes be subject to error. Our approach to tackle the problem involves assuming that the observed classifications are imperfect measures of an underlying hidden colonisation process.

The observed data for an individual c in pen p are collected in prescheduled observation times, which we denote by $O^{c,p} = \left\{O_g^{c,p} \cup O_\pm^{c,p}\right\} \subseteq \mathcal{T}^{c,p}$, where $O_g^{c,p}$ is defined as the set of genotyped observation times and $O_\pm^{c,p} = O^{c,p} \setminus O_g^{c,p}$ are the times where no genotyping was done. Moreover, let $U^{c,p} = \mathcal{T}^{c,p} \setminus O^{c,p}$ denote the times that the individual was not tested. Let $R_t^{[c,p]}$ and $F_t^{[c,p]}$ denote the outcome of the RAMS and faecal test, respectively, recorded at time $t \in O^{c,p}$. When $t \in O_\pm^{c,p}$ a test result is classified as negative, denoted by 0, or positive, denoted by $+$. When a positive test is genotyped at $t \in O_g^{c,p}$, then we can further characterise the test as s-genotype positive (when a type s is detected), denoted by $s \in \{1, 2, \ldots, n_g\}$.

We assume that the RAMS and faecal tests are independent conditional on the carriage status of the individual. Moreover, the observed states $R_t^{[c,p]}$ and $F_t^{[c,p]}$ are generated conditional on the carriage state $X_t^{[c,p]}$ according to a misspecification matrix $\mathbf{E}^R$ and $\mathbf{E}^F$, respectively, with elements $e_{r,s}^R = \mathbb{P}\left(R_t^{[c,p]} = s \middle| X_t^{[c,p]} = r\right)$ and $e_{r,s}^F = \mathbb{P}\left(F_t^{[c,p]} = s \middle| X_t^{[c,p]} = r\right)$.

We distinguish two cases: tests not chosen to be genotyped and tests that were genotyped. For the case where a positive RAMS sample was not chosen to be genotyped we assume that both the RAMS and the faecal tests have 100% specificity (i.e. it is not possible to test positive when the true carriage status is non-carrier) and so the observation matrix $\mathbf{E}^{R_\pm} = \left\{e_{r,s}^{R_\pm}\right\}_{r \in \{0,1,\ldots,n_g\};\, s \in \{0,+\}}$ is given by:

$$
\begin{array}{cc} & \begin{array}{cc} \multicolumn{2}{c}{\text{observed state}} \\ 0 & + \end{array} \\
\begin{array}{c} \\ 0 \\ 1 \\ \text{hidden} \\ \text{true state} \\ \vdots \\ n_g \end{array} &
\begin{bmatrix} 1 & 0 \\ 1-\theta_R & \theta_R \\ \vdots & \vdots \\ 1-\theta_R & \theta_R \end{bmatrix}
\end{array}
\tag{2.4}
$$

and similarly, for the faecal test with θ_R replaced with θ_F. Here, the test sensitivities are denoted by $\theta_R = \mathbb{P}\left(R_t^{[c,p]} = + \middle| X_t^{[c,p]} = r\right)$ and $\theta_F = \mathbb{P}\left(F_t^{[c,p]} = + \middle| X_t^{[c,p]} = r\right)$, for $r = 1, 2, \ldots, n_g$.

For a positive sample that was genotyped we introduce additional parameters θ_C, θ_S and θ_P to allow for the possibility of genotype misspecification. The parameters have the following interpretations. Given that a test is found positive, θ_C denotes the probability

of correctly identifying a common genotype $\{1, 2, \ldots, n_g - 1\}$, θ_S is the probability of misclassifying a common genotype with a different common genotype, and θ_P the probability that a genotype of pooled type n_g is classified as a common genotype. We assume that these probabilities are the same for both the RAMS and faecal tests. More specifically, the matrix of classification probabilities for the RAMS test, $\mathbf{E}^{R_g}$, is a $(n_g + 1) \times (n_g + 1)$ matrix of the form:

$$
\begin{array}{cc}
 & \\
 & \\
\text{hidden true state} & \begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \\ n_g - 2 \\ n_g - 1 \\ n_g \text{ (Pooled)} \end{array}
\end{array}
\begin{array}{c}
\begin{array}{ccccccc} \multicolumn{7}{c}{\text{observed state}} \\ 0 & 1 & 2 & 3 & \cdots & n_g - 1 & n_g \text{ (Pooled)} \end{array} \\
\left[\begin{array}{ccccccc}
1 & 0 & 0 & 0 & \cdots & 0 & 0 \\
1-\theta_R & \theta_C\,\theta_R & \dfrac{\theta_S\,\theta_R}{n_g-2} & \dfrac{\theta_S\,\theta_R}{n_g-2} & \cdots & \dfrac{\theta_S\,\theta_R}{n_g-2} & (1-\theta_C-\theta_S)\,\theta_R \\
1-\theta_R & \dfrac{\theta_S\,\theta_R}{n_g-2} & \theta_C\,\theta_R & \dfrac{\theta_S\,\theta_R}{n_g-2} & \cdots & \dfrac{\theta_S\,\theta_R}{n_g-2} & (1-\theta_C-\theta_S)\,\theta_R \\
\vdots & \vdots & & \ddots & & & \vdots \\
1-\theta_R & \dfrac{\theta_S\,\theta_R}{n_g-2} & \cdots & \dfrac{\theta_S\,\theta_R}{n_g-2} & \theta_C\,\theta_R & \dfrac{\theta_S\,\theta_R}{n_g-2} & (1-\theta_C-\theta_S)\,\theta_R \\
1-\theta_R & \dfrac{\theta_S\,\theta_R}{n_g-2} & \cdots & \dfrac{\theta_S\,\theta_R}{n_g-2} & \dfrac{\theta_S\,\theta_R}{n_g-2} & \theta_C\,\theta_R & (1-\theta_C-\theta_S)\,\theta_R \\
1-\theta_R & \dfrac{\theta_P\,\theta_R}{n_g-1} & \cdots & \dfrac{\theta_P\,\theta_R}{n_g-1} & \dfrac{\theta_P\,\theta_R}{n_g-1} & \dfrac{\theta_P\,\theta_R}{n_g-1} & (1-\theta_P)\,\theta_R
\end{array}\right]
\end{array}
\tag{2.5}
$$

such that, for all $r \neq 0$, the probabilities $e^{R_g}_{r,0} = \mathbb{P}\left(R_t^{[c,p]} = 0 \middle| X_t^{[c,p]} = r\right) = 1 - \theta_R$ and $\sum_{s=1}^{n_g} e^{R_g}_{r,s} = \theta_R$. The misclassification matrix for the faecal test is defined similarly replacing θ_R with θ_F in matrix (2.5).

2.4. *Model fitting.* The approach adopted in this paper uses Bayesian data augmentation methods, in which the unobserved carriage states are treated as additional parameters and are imputed from the data. This is facilitated by the use of MCMC algorithms. Let $\mathbf{X}_t^{[1:C,p]}$ be the vector of the hidden carriage states for individuals $1, 2, \ldots, C$ in pen p at time t and $\mathbf{X} = \left\{\mathbf{X}_t^{[1:C,p]}\right\}_{p\in\{1,2,\ldots,P\};\, t\in\mathcal{T}^{c,p}}$ be the whole hidden state process (i.e. over all individuals in all pens and for all time points). Similarly, the observed longitudinal data comprises RAMS and faecal test results, denoted by $\mathbf{R} = \left\{\mathbf{R}_t^{[1:C,p]}\right\}_{p\in\{1,2,\ldots,P\};\, t\in\mathcal{T}^{c,p}}$ and $\mathbf{F} = \left\{\mathbf{F}_t^{[1:C,p]}\right\}_{p\in\{1,2,\ldots,P\};\, t\in\mathcal{T}^{c,p}}$ respectively. We use the notation $\boldsymbol{\theta} = (\theta_R, \theta_F, \theta_C, \theta_S, \theta_P)$ for the observation parameters, and $\boldsymbol{\phi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\nu}, \gamma, \delta)$ for the transmission parameters, where $\boldsymbol{\alpha} = \{\alpha_s\}_{s=1}^{n_g}$, $\boldsymbol{\beta} = \{\beta_s\}_{s=1}^{n_g}$, $\boldsymbol{\mu} = \{\mu_s\}_{s=1}^{n_g}$ and $\boldsymbol{\nu} = \{\nu_s\}_{s=1}^{n_g}$.

The Bayesian approach requires the specification of the prior distributions over the model parameters $\boldsymbol{\psi} = (\boldsymbol{\phi}, \boldsymbol{\theta})$, $\pi(\boldsymbol{\psi})$. For the genotype-specific external colonisation rates, the within-pen colonisation rates and the clearance rates, we assigned weakly-informative

univariate Exponential priors each with mean 1. The priors for δ and γ are also assumed to be Exponential with rate parameter ln(2), reflecting equal prior probabilities for these parameters to be less or more than one. We assume Beta(0.5, 0.5) prior distributions for the sensitivity parameters θ_R, θ_F and θ_P, which is the Jeffreys' prior (Jeffreys, 1961). For the remaining sensitivity parameters we assume a minimally informative Dirichlet prior distribution (Kelly and Atwood, 2011) with $E(\theta_C) = 0.9$, that is, $(\theta_C, \theta_S, 1-\theta_C-\theta_S) \sim \text{Dirichlet}(4.5, 0.25, 0.25)$. Finally, for the probabilities of carriage at the beginning of the study we use a Jeffreys' Dirichlet distribution with all (n_g+1) parameters set to 0.5.

Combining the complete data likelihood with the prior allows us to formulate the joint posterior distribution of the hidden carriage states (unobserved data) and the model parameters which can be factorised as:

$$
\begin{aligned}
\pi(\mathbf{X}, \boldsymbol{\phi}, \boldsymbol{\theta} \mid \mathbf{R}, \mathbf{F}) &\propto \pi(\mathbf{R}, \mathbf{F} \mid \mathbf{X}, \boldsymbol{\theta})\, \pi(\mathbf{X} \mid \boldsymbol{\phi})\, \pi(\boldsymbol{\psi}) \\
&= \prod_{p=1}^{P} \prod_{c=1}^{C} \Bigg[\prod_{r=0}^{n_g} \prod_{s \in \mathcal{X}_g} \prod_{t \in O_g^{c,p}} \left[\left(e_{r,s}^{R_g}\right)^{\mathbb{1}\left\{X_t^{[c,\,p]}=r, R_t^{[c,\,p]}=s\right\}} \left(e_{r,s}^{F_g}\right)^{\mathbb{1}\left\{X_t^{[c,\,p]}=r, F_t^{[c,\,p]}=s\right\}} \right] \\
&\quad \times \prod_{r=0}^{n_g} \prod_{s \in \{0,+\}} \prod_{t \in O_{\pm}^{c,p}} \left[\left(e_{r,s}^{R_{\pm}}\right)^{\mathbb{1}\left\{X_t^{[c,\,p]}=r, R_t^{[c,\,p]}=s\right\}} \left(e_{r,s}^{F_{\pm}}\right)^{\mathbb{1}\left\{X_t^{[c,\,p]}=r, F_t^{[c,\,p]}=s\right\}} \right] \\
&\quad \times \prod_{s=0}^{n_g} \nu_s^{\mathbb{1}\left\{X_1^{[c,\,p]}=s\right\}} \times \prod_{r=0}^{n_g} \prod_{s=0}^{n_g} \prod_{t \in \mathcal{T}^{c,p} \setminus \{1\}} \left[\left(m_{r,s}^{p}(t)\right)^{\mathbb{1}\left\{X_{t-1}^{[c,\,p]}=r, X_t^{[c,\,p]}=s\right\}} \right] \Bigg] \\
&\quad \times \pi(\boldsymbol{\psi}), \qquad (2.6)
\end{aligned}
$$

where $\mathbb{1}_A$ is the indicator function of event A. The factorisation in Equation (2.6) is based on the assumption that conditionally on the model parameters, the carriage process is assumed to be independent across pens.

Sampling from the posterior distribution is done by constructing an MCMC algorithm that employs both Gibbs (Geman and Geman, 1984) and Hamiltonian Monte Carlo (HMC; Neal, 2011) updates. The main emphasis is on sampling the hidden carriage process $\mathbf{X}$, which was done by using a Gibbs step via the recent individual-Forward Filtering Backward Sampling (iFFBS) algorithm by Touloupou, Finkenstädt and Spencer (2019). A point which is worth emphasising is that iFFBS method has made it computationally feasible to fit such a complex model, which we believe would not have been able to be fitted using any other existing methodologies. For example, the vanilla Forward Filtering Backward Sampling (Carter and Kohn, 1994) method in our setting, with 8 genotypes and 8 animals per pen, is computationally infeasible since the transition matrix has $8^{2\times 8} = 2.81475 \times 10^{14}$ elements. The initial probability parameters $\boldsymbol{\nu}$ and the observation parameters $\boldsymbol{\theta}$ are updated using Gibbs updates. The remaining parameters are updated jointly using an HMC algorithm, which requires the partial derivatives for these parameters found in Equations (A.1)-(A.4) of the Supplementary Material.

The inference method for the proposed partially observed multi-strain model was implemented in the R programming language (R Core Team, 2019) using the recently developed

package epiPOMS (Touloupou and Spencer, 2020).

3. Results.

3.1. *Simulated data analysis.* The performance of our Bayesian approach is evaluated via simulation studies under different settings. In the first setting, we simulate data with the same structure as the observed data, and in the second setting we investigate the effect of the total number of samples that are genotyped per pen. A full description of the analysis can be found at Supplementary Material in Section B and here we summarize the key results from the simulations.

Priors specifications are identical to the ones defined in Section 2.4. In terms of parameter estimation, all the simulation studies show that the method was able to identify and provide estimates in the sense that in all settings the 90% quantile intervals of the posterior medians, over 50 simulated datasets, contained the true parameter values that were used to generate the data, see Figures B.1 and B.4 of the Supplementary Material for each of the two settings respectively. Another important task in the estimation procedure is recovering the hidden carriage process. Using the augmented states of carriage in each MCMC iteration, one can plot the Receiver Operating Characteristic (ROC) curve for each genotype, shown in Figures B.3 and B.5 of the Supplementary Material. These figures indicate that the method reproduced the incidence of colonisation with high accuracy, since in all genotypes the ROC curve is located close to the top left corner.

Finally, in order to investigate the performance of our approach subject to the amount of genotyped samples per pen, we simulated different datasets representing situations with sparse, moderate and dense genotyping. We showed that the performance of our method depends on the amount of genotyped samples: as the total number of genotypes increases, the accuracy of the estimate also increases for all model parameters (see Figure B.4 of the Supplementary Material). We can reconstruct the genetic type of infection with high probability from surprisingly few typed observations, as can be seen from the ROC curves in Figure B.5 of the Supplementary Material and the area under the curve (AUC) in Table 3. More specifically, the median AUC value is found to be above 0.95 in all settings. An AUC above 95% indicates that our method gives a higher posterior probability to positive samples than negative samples more than 95% of the time, even with only 5 typed samples per pen.

3.2. *Data analysis.* In this section, we apply our Bayesian data augmentation approach to the observed *E. coli* O157:H7 data described in Section 2.1. Our goals are to obtain estimates for the epidemiologically important parameters and to investigate possible differences between genotypes in carriage colonisation and clearance.

We ran the MCMC for 35,000 iterations, discarding the first 10,000 as a burn-in and saved every 5 iterations to obtain 5,000 samples from the posterior. We used the same priors as in the simulation studies of Section 3.1. Convergence was assessed by visual inspection of posterior trace plots for all 39 model parameters, shown in Figure C.2 of the Supplementary Material. We also checked that estimates were robust to a change in the initial values by running 3 chains starting from diverse values. Convergence of the hidden state process was also visually assessed. In particular, trace plots of two different

TABLE 3
Medians of the area under the ROC curve over 50 simulated data sets with varying levels of genotyping.

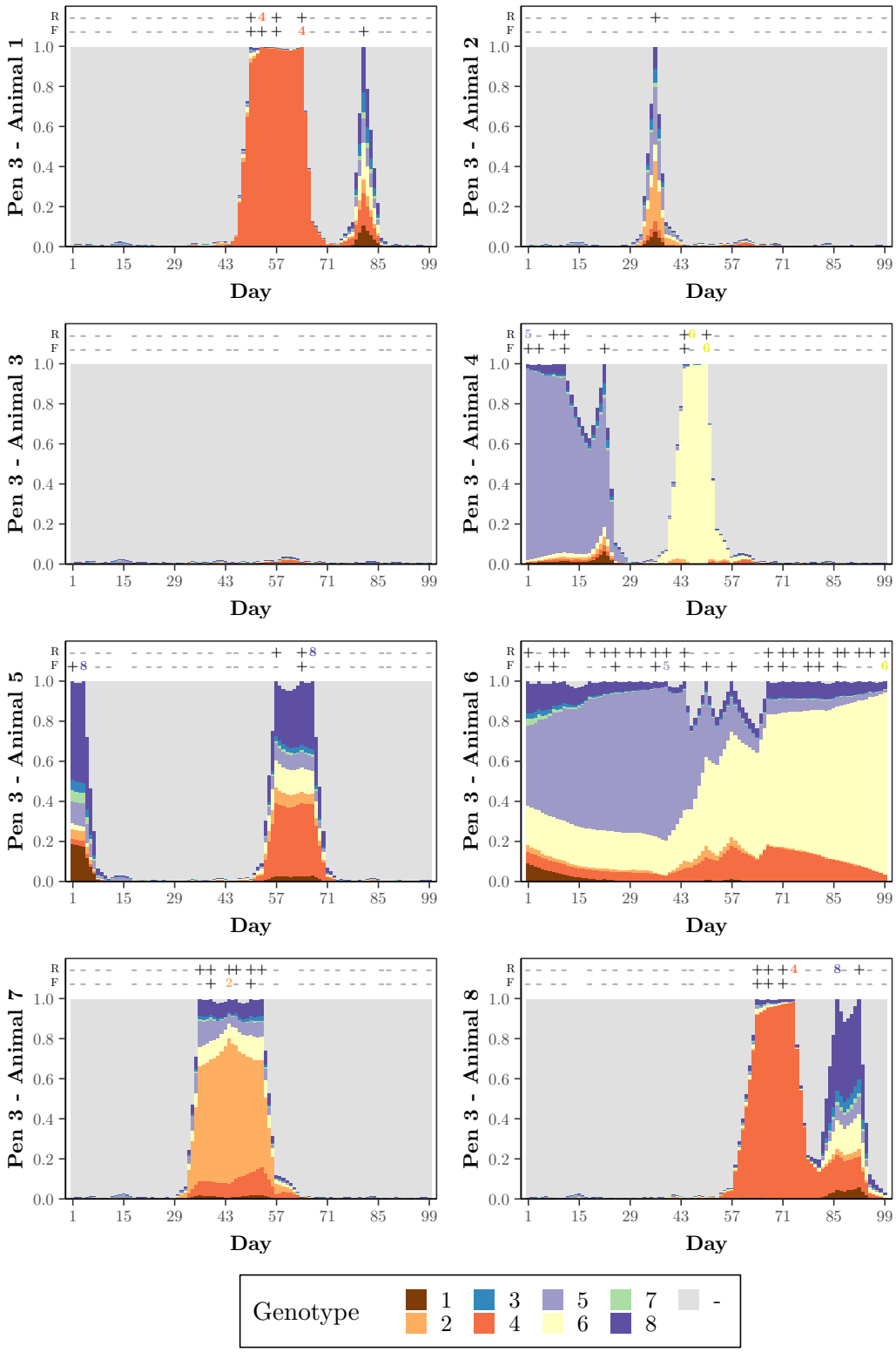
Genotype	Total number of genotyped samples per pen				
	5	10	15	20	All
D (1)	0.977	0.982	0.985	0.988	0.993
J (2)	0.970	0.979	0.983	0.983	0.991
X (3)	0.972	0.983	0.985	0.988	0.992
b (4)	0.967	0.974	0.981	0.981	0.991
d (5)	0.970	0.982	0.985	0.987	0.992
f (6)	0.969	0.977	0.985	0.983	0.991
l (7)	0.967	0.977	0.984	0.986	0.992
Pooled (8)	0.952	0.969	0.971	0.980	0.986

summary statistics are provided: the first is the total number of individuals in the population having a specific augmented state of carriage over the entire study period, whilst the second summary statistic is the total number of genotype-specific transitions in the augmented carriage process over all individuals. The trace plots are shown in Figure C.3 of the supplementary material and demonstrate very good mixing of the chains, which appear to reach their stationary distribution.

In addition, estimated posterior probabilities of colonisation by a specific genotype are shown for individuals in pen 3 in Figure 4. Even though samples were taken only twice per week, the algorithm provides probabilities of colonisation for every day of the study period. Moreover, our method predicts the genotype of infection for an individual whose samples were not genotyped. This is achieved by borrowing information from the other individuals within the pen or from the individual itself in a nearby time point. For example, we observe that for individual 6 the method assigns a non-zero probability of the individual to be colonised by type 6 during the whole study period, even though the faecal sample at day 39 identified type 5. This happens because a few days earlier and after there were individuals that had a positive type 6 result and therefore the method allows for the possibility that individual 6 was also colonised by the same type. When sequences of positive test results separated by a negative result occur, the method can quantify the probability of re-infection versus a false negative test. An example is individual 4 where we can see a spike of grey (representing the individual becoming uncolonised) immediately after day 15. The convergence and mixing properties of the augmented unobserved carriage states were further evaluated by running the MCMC algorithm for 40000 iterations and splitting the last 30000 into 3 batches of equal size (where a thinning of 5 was applied to the samples). The posterior probabilities of colonisation for each individual, were estimated using each batch of 10000 samples. Results are shown in Figure C.4 of the supplementary material for individual 6 in pen 3. We see that all plots provide identical results, indicating no evidence of non-convergence or poor mixing.

Tables 4 shows the posterior median estimates of the transmission parameters, along with 90% credible intervals. We see that the external colonisation rate for the pooled group, type 8, is uniformly higher than the rest of the types. This is due to the fact that

FIG 4. *Posterior probability of colonisation over time with separate plots for each individual within Pen 3 in the E. coli O157:H7 data. In each figure the top panel contains the observed test results, where the first line represents the outcome of RAMS samples and the second line represents the outcome of faecal samples. "-" indicates negative sample, "+" indicates that the sample was positive but not chosen for genotyping, otherwise, the genotype name is given.*



α_8 accounts for all acquisitions of the 41 genotypes in the group. The lowest external colonisation rate belongs to genotype 7, following by 4, 6 and 2. However, most of the differences are not significant, as suggested by the overlap of the credible intervals for all parameters. Moreover, we see that the sum of the posterior medians of the eight genotype-specific external colonisation rates ($\sum_{s=1}^{n_g} \alpha_s = 0.008$ per day) is in close agreement with the external colonisation rate estimated in a non-genotype specific analysis in Spencer et al. (2015) ($\alpha = 0.009$ per day).

The posterior median for the within-pen colonisation rate was almost 4.5 times higher for genotype 7 (0.0157 per day) compared to genotype 6 (0.0035 per day) with non-overlapping 90% credible intervals, which suggests that there are differences in the within-pen colonisation rates between the studied genotypes. The estimates of within-pen colonisation rate for the remaining genotypes were estimated to be between these two values. Results suggest no significant differences in durations of carriage ($1/\mu_s$) between genotypes. In particular, in Table 4 we see that all parameters have overlapping credible intervals.

As a general finding, we observe that genotype 7 appears to have the highest within-pen but the lowest external colonisation rate suggesting that it is mainly transmitted through contact between animals in the same pen. Similarities between colonisation rates are found for genotypes 1 with 3 and also 4 with 6 (Table 4). The latter genotypes (4 and 6) are the least prevalent (Figure C.1 in Supplementary Material, as calculated from the latent carriage process) which explains their low within-pen and external transmission rates.

The relative colonisation rate γ in smaller versus bigger pens was estimated as 1.499 with a 90% credible interval of [0.826, 2.285], suggesting higher rates of infection when animals are kept closer together. However, since the interval contains 1 this difference is not significant. This finding is consistent with Spencer et al. (2015), where a similar relative difference between pens was obtained. The posterior median relative colonisation rate in a carrier versus non-carrier individual was 0.842 (90% credible interval [0.001, 1.806]) which indicates that individuals colonised by one genotype were less likely to acquire a new infection with another genotype. However, the 90% credible interval contains 1 indicating no significant effect.

Table 5 shows posterior summaries for the observation parameters. As in a previous analysis (Spencer et al., 2015), we find that the test sensitivities θ_R and θ_F are 0.76 and 0.46, respectively. Additionally, the model estimates that 81.6% of the observed common genotypes are correctly classified as the right type, 1.2% are misclassified as another common type and the remaining 17.2% are misclassified as type 8. Finally, we estimate that 98% of the observed 8 genotypes are correctly classified as type 8.

To explore the effect of our prior specifications, we perform a sensitivity analysis using different hyperparameter values each time. Results are shown in Supplementary Material in Figure C.5 for parameters γ and δ, Figure C.6 for the observation parameters, and Figure C.7 for the transmission and clearance rate parameters. The posterior distributions of the within-pen and external transmission rates, as well as the clearance rates remained unchanged. No major change is observed in the posterior median and quantiles of the observation parameters when replacing the minimally informative Dirichlet prior with one of the other two non informative alternatives. Finally, the use of an $\text{Exp}\,(2\ln(2))$ prior for parameters γ and δ leads to a decrease in their posterior median, compared to the use of

TABLE 4

Estimates of genotype-specific transmission model parameters among cattle in the E. coli O157:H7 data: the posterior median of the parameter and the 90% credible interval within parentheses. Estimates are multiplied by 100.

Genotype (s)	**Transmission model parameter** ($\times 100$)			
	Initial probability of carriage (ν_s)	External rate (α_s, day^{-1})	Within-pen rate (β_s, day^{-1})	Clearance rate (μ_s, day^{-1})
D (1)	2.909 (0.314, 5.814)	0.123 (0.050, 0.204)	0.989 (0.367, 1.693)	16.104 (9.713, 23.002)
J (2)	0.556 (0.000, 2.455)	0.080 (0.024, 0.152)	1.222 (0.290, 2.411)	17.164 (8.574, 27.164)
X (3)	0.686 (0.000, 2.484)	0.122 (0.054, 0.203)	1.093 (0.473, 1.834)	13.310 (8.460, 18.431)
b (4)	0.261 (0.000, 1.492)	0.058 (0.011, 0.110)	0.620 (0.003, 1.259)	9.789 (3.268, 17.734)
d (5)	1.628 (0.000, 3.896)	0.146 (0.063, 0.231)	0.693 (0.276, 1.169)	9.964 (6.080, 14.202)
f (6)	0.314 (0.000, 1.667)	0.059 (0.013, 0.118)	0.347 (0.000, 0.845)	6.853 (0.743, 16.849)
l (7)	0.955 (0.000, 2.601)	0.046 (0.009, 0.094)	1.571 (0.901, 2.345)	11.767 (7.268, 17.091)
Pooled (8)	2.119 (0.000, 5.443)	0.192 (0.086, 0.314)	0.723 (0.274, 1.186)	9.501 (6.081, 13.002)

TABLE 5

Estimates of observation model parameters among cattle in the E. coli O157:H7 data: the posterior median of the parameter and the 90% credible interval within parentheses. Estimates are given as percentages.

Observation model parameter (%)				
θ_R	θ_F	θ_C	θ_S	$1-\theta_P$
76.4 (72.6, 80.0)	45.7 (42.0, 49.2)	81.6 (74.6, 88.7)	1.2 (0.0, 4.4)	97.9 (90.8, 100)

an uninformative prior Exp (0.01). A possible explanation is that our data are only weakly informative due to the relatively small number of type-specific transitions $r \mapsto s$, where $r, s \neq 0$ (posterior median 29, 90% credible interval [1, 59]).

For comparison, we also fitted simpler models to the data, namely a model where a common clearance rate is assumed for the common genotypes, a model in which we set $\delta = 0$ and so carriers cannot be infected until they have cleared their current strain, and a model with $\gamma = 1$ in which there is no difference in colonisation rates between individuals housed in small and large pens. Posterior distributions for the parameters of interest are shown in Figure C.8 of the Supplementary Material. As expected, when a common clearance rate is assumed, the new estimate is approximately equal to the average of the full

model estimates for common genotypes but less associated variability. This leads to minor changes in external and within-pen colonisation rates. When $\gamma = 1$, no profound change is observed in the posterior distribution of the model parameters, except for the within-pen colonisation rates where estimates are higher compared to the estimates obtained using the other models. This is expected, since now the $\boldsymbol{\beta}$ parameters must take values that account for the number of infections occurring within pens with smaller area, for which γ parameter accounts in the remaining models. Finally, setting $\delta = 0$ has small effects on the posterior estimates of the parameters. In particular, we observe a higher degree of genotype misclassification, as can be seen by the slightly lower and higher estimates of parameters θ_C and θ_S, respectively. This happens because the model does not allow for genotype changes, and therefore, when two consecutive observations of different genotypes occur, they are either attributed to type misspecification or to true conversions. However, due to the very low number of genotype-to-genotype transitions that are estimated in the full model, we find only little difference in the estimates.

To conclude, we highlight that the typing information gives a better understanding of who infected whom within a pen. In particular, the multi-strain model allows us to say with greater certainty which infections were transmitted within the pen and which were acquired from outside the pen. Moreover, the new insights we uncover regarding the genotype-specific properties, related to transmission, as well as competition between different genotypes and genotype misspecification, represent important aspects of the analysis of host-pathogen interactions which would have been overlooked if we had considered the data with the strains pooled together.

4. Discussion. In this paper, we have developed a model for analysing longitudinal carriage studies with multiple strains. Our model extends existing methodologies (Cauchemez et al., 2006; Melegaro et al., 2007; Erästö, Hoti and Auranen, 2012; Numminen et al., 2013) by allowing imperfect classification, that is, that the true carriage states can be falsely recorded as non-carrier or misclassified as another genotype. Furthermore, it gains flexibility by allowing non-typed samples to be classified as any of the studied genotypes rather than pooling them into the pooled group, as is assumed by the majority of the aforementioned models. We assume that during the study no evolution occurs that results in a change in the PFGE banding pattern, and so our approach is unsuitable for high resolution genotying data unless some clustering has first been applied to reduce the resolution of the data.

Although our method was motivated by a study of repeated observations of *E. coli* O157:H7 colonisation, it can be applied with minor modifications to other infectious diseases. Important examples might include the transmission of strains of *Staphylococcus aureus* within healthcare settings, where our modelling approach could make inferences about the differences in transmissibility and competition between methicillin resistant and non-methicillin resistant strains. Dengue virus has five known types (Normile, 2013), many of which co-circulate (Messina et al., 2014) and repeated infection is thought to bring more serious consequences for individuals. Improving understanding of the interaction between competing strains of influenza A could help in vaccine design and pandemic preparedness. In these important human examples, pens would be replaced by households and additional parameters would be needed to describe contacts between households.

The algorithm proposed in Touloupou, Finkenstädt and Spencer (2019) has made it computationally feasible to fit the multi-strain epidemic model described in this paper. An advantage of this algorithm compared to previous approaches for fitting multi-strain models is that it can be efficiently applied with several genotypes and can reduce the correlation between posterior samples by updating the entire hidden carriage process at each iteration. Moreover, the computation time required for the algorithm can scale linearly with population size and in the cube of number of different genotypes, making it applicable to much larger datasets.

Simulations demonstrated that the algorithm accurately estimated the model parameters and successfully reproduced the incidence of colonisation. A sensitivity analysis was conducted to explore different genotyping strategies, and indicated that a surprisingly small number of genotyped samples were required to successfully recover the strain dynamics. This suggests that in future studies, a simulation study like the one described in supplementary information can be used to design more efficient genotyping schemes, based on parameter estimates from an initial study.

Application of our method to a longitudinal study of *E. coli* O157:H7 in feedlot cattle has given us valuable insights into the multi-strain dynamics within herds. The analysis has been implemented in the R package epiPOMS by Touloupou and Spencer (2020). Results provided evidence for between-genotype competition, as the relative colonisation rate for carriers versus non-carriers was estimated to be 0.85. Smaller pens with higher stocking densities were more susceptible to within-pen colonisation compared to larger pens, as suggested by a relative transmission rate of 1.5. However, both credible intervals contain the value of one, indicating no clear evidence of a difference.

Differences between genotypes were detected with respect to the external rate of colonisation. In particular, we found that genotype 8 (pooled group) has the highest rate while genotype 7 has the lowest. This is expected given that type 8 represents a total of 41 genotypes in the data. For the within-pen colonisation rate we found significant differences between genotypes 6 and 7, the latter being 4.5 times the former. Similarities were observed for genotypes 1 and 3 in terms of both external and within-pen colonisation rates. Genotypes 4 and 6 shared a low external colonisation rate as well as within-pen rate. Clearance rates were found to be relatively homogeneous between genotypes, in the sense that posterior credible intervals for these parameters were overlapping.

A significant merit of our approach is that it allows for the presence of imperfect diagnostic tests. We estimated that in the real data the sensitivity of the faecal test was as low as 46% and the sensitivity of the RAMS test was close to 76%. In addition, we estimated that only 82% of the observed common genotypes were correctly identified. These findings highlight the importance of allowing for imperfect tests within the model, an assumption which has been ignored in several previous epidemiological studies.

Following previous work, our model treats different genotypes that appear rarely in the observed data as a single pooled group. Even though this assumption is unrealistic, we believe that this does not impair our inferences for the parameters relating to common genotypes, and we don't provide any epidemiological interpretation for the within-pen transmission parameter of the pooled group. Further, we allow the pooled group to have a different probability of genotype misclassification to the other groups.

One potential limitation of our model is that we currently do not allow for co-infection, that is, an individual carrying more than one genotype at a time. This assumption may be reasonable for the current *E. coli* O157 dataset, as there was infrequent evidence of co-infection in this study. However co-infection models may be applicable for other epidemiological studies of microbial transmission where mixed infections are more common. Nevertheless, one can envisage that our model can be extended to allow for colonisation by all pairwise combinations of single carriage states and the same algorithm can be used for posterior inferences. Another extension of our model that one may consider is accounting for between-pen interactions; this can be achieved by adding an extra between-pen transmission parameter as was done in Touloupou (2016) for a single-strain model. However, on account of the findings there, which suggested no interaction between pens in this particular study, we chose not to consider this extension here.

In conclusion we've used state-of-the-art techniques from Bayesian inference, such as the individual forward filtering backward sampling algorithm for coupled hidden Markov models and Hamiltonian MCMC, to make inferences for the transmission dynamics of a multi-strain epidemic model. Furthermore our model contains some novel features, such as the ability to account for strains to be misclassified in the data. Altogether, this has generated novel insights into the epidemiology of infection for *Escherichia coli* O157:H7 in feedlot cattle that would not have been possible using previous approaches.

Acknowledgements. PT was supported by a University of Warwick Department of Statistics PhD scholarship. SEFS gratefully acknowledges funding by MRC grant MR/P026400/1 and EPSRC grant EP/R018561/1. The original data set was generated from work funded by the Beef Checkoff, with support from National Institutes of Health Public Health Service grants U54-AI-57141, P20-RR16454 and P20-RR15587, and by Agriculture and Food Research Initiative competitive grant no. 2010-04487 from the USDA National Institute of Food and Agriculture. The authors would like to thank Rowland Cobbold for sharing the data from the original study.

Supplementary Materials. This article is accompanied by two supplementary files:

Supplement A: Appendices (.pdf). In the appendices, we provide further details on algorithmic implementation. Moreover, we show results of the simulations studies on synthetic data to investigate our estimation procedure. Additional plots and results of the real data analysis can also be found in the appendices.

Supplement B: epiPOMS R package (.tar.gz). The R package epiPOMS provides tools for Bayesian inference on epidemiological data using partially observed multi-strain (POMS) epidemic models, focusing on applications where observations are gathered longitudinally and the population under investigation is organised in small groups. The *E. coli* 0157:H7 multi-strain dataset is provided there as an example.

References.

Anderson, R. M. and May, R. M. (1991). *Infectious diseases of humans: Dynamics and control*. Oxford University Press.

Auranen, K., Arjas, E., Leino, T. and Takala, A. K. (2000). Transmission of *pneumococcal* carriage in families: A latent Markov process model for binary longitudinal data. *Journal of the American Statistical Association* **95** 1044-1053.

Becker, N. G. (1989). *Analysis of infectious disease data. Chapman and Hall/CRC Monographs on Statistics and Applied Probability.* Taylor and Francis.

Becker, N. G. (1997). Uses of the EM algorithm in the analysis of data on HIV/AIDS and other infectious diseases. *Statistical Methods in Medical Research* **6** 24–37.

Becker, N. G. and Britton, T. (1999). Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61** 287–307.

Becker, N. G. and Hasofer, A. M. (1997). Estimation in epidemics with incomplete observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59** 415–429.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81** 541–553.

Cauchemez, S., Temime, L., Valleron, A.-J., Varon, E., Thomas, G., Guillemot, D. and Boëlle, P.-Y. (2006). *S. pneumoniae* transmission according to inclusion in conjugate vaccines: Bayesian analysis of a longitudinal follow-up in schools. *BMC Infectious Diseases* **6** 1-10.

Clancy, D. and O'Neill, P. D. (2007). Exact Bayesian inference and model selection for stochastic models of epidemics among a community of households. *Scandinavian Journal of Statistics* **34** 259–274.

Cobbold, R. N., Hancock, D. D., Rice, D. H., Berg, J., Stilborn, R., Hovde, C. J. and Besser, T. E. (2007). Rectoanal junction colonization of feedlot cattle by *Escherichia coli* O157:H7 and its association with supershedders and excretion dynamics. *Applied and Environmental Microbiology* **73** 1563-1568.

Daley, D. J. and Gani, J. (2001). *Epidemic modelling: An introduction.* Cambridge University Press.

Diekmann, O., Heesterbeek, H. and Britton, T. (2012). *Mathematical tools for understanding infectious disease dynamics.* Princeton University Press.

Dong, W., Pentland, A. and Heller, K. A. (2012). Graph-coupled HMMs for modeling the spread of infection. *arXiv preprint arXiv:1210.4864.*

Erästö, P., Hoti, F. and Auranen, K. (2012). Modeling transmission of multitype infectious agents: Application to carriage of *Streptococcus pneumoniae*. *Statistics in Medicine* **31** 1450-1463.

Fearnhead, P. and Meligkotsidou, L. (2004). Exact filtering for partially observed continuous time models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66** 771–789.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721-741.

Gibson, G. J. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology* **15** 19–40.

Ionides, E. L., Bretó, C. and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* **103** 18438-18443.

Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S. and King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences* **112** 719–724.

Jeffreys, H. (1961). *Theory of probability.* Oxford: Clarendon Press.

Jewell, C. P., Kypraios, T., Neal, P. and Roberts, G. O. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis* **4** 465–496.

Kelly, D. and Atwood, C. (2011). Finding a minimally informative Dirichlet prior distribution using least squares. *Reliability Engineering and System Safety* **96** 398–402.

Kypraios, T., Neal, P. and Prangle, D. (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation. *Mathematical Biosciences* **287** 42–53.

Kypraios, T., ONeill, P. D. et al. (2018). Bayesian nonparametrics for stochastic epidemic models. *Statistical Science* **33** 44–56.

Kypraios, T., O'Neill, P. D., Huang, S. S., Rifas-Shiman, S. L. and Cooper, B. S. (2010). Assessing the role of undetected colonization and isolation precautions in reducing Methicillin-Resistant *Staphylococcus aureus* transmission in intensive care units. *BMC Infectious Diseases* **10** 1–10.

Lord, C., Barnard, B., Day, K., Hargrove, J., McNamara, J., Paul, R., Trenholme, K. and Woolhouse, M. (1999). Aggregation and distribution of strains in microparasites. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **354** 799–807.

McKinley, T., Cook, A. R. and Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics* **5** 24.

McKinley, T. J., Ross, J. V., Deardon, R. and Cook, A. R. (2014). Simulation-based Bayesian inference for epidemic models. *Computational Statistics and Data Analysis* **71** 434–447.

Melegaro, A., Choi, Y., Pebody, R. and Gay, N. (2007). *Pneumococcal* carriage in United Kingdom families: Estimating serotype-specific transmission parameters from longitudinal data. *American Journal of Epidemiology* **166** 228-235.

Messina, J. P., Brady, O. J., Scott, T. W., Zou, C., Pigott, D. M., Duda, K. A., Bhatt, S., Katzelnick, L., Howes, R. E., Battle, K. E. et al. (2014). Global spread of dengue virus types: mapping the 70 year history. *Trends in microbiology* **22** 138–146.

Morton, A. and Finkenstädt, B. F. (2005). Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54** 575–594.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Mote Carlo* (S. Brooks, A. Gelman, G. L. Jones and X. Meng, eds.) 5, 113–162. Chapman & Hall/CRC.

Neal, P. (2012). Efficient likelihood-free Bayesian computation for household epidemics. *Statistics and Computing* **22** 1239–1256.

Normile, D. (2013). Surprising new dengue virus throws a spanner in disease control efforts.

Numminen, E., Cheng, L., Gyllenberg, M. and Corander, J. (2013). Estimating the transmission dynamics of *Streptococcus pneumoniae* from strain prevalence data. *Biometrics* **69** 748-757.

O'Neill, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences* **180** 103–114.

O'Neill, P. D. and Becker, N. G. (2001). Inference for an epidemic when susceptibility varies. *Biostatistics* **2** 99-108.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **162** 121-129.

Rida, W. N. (1991). Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *Journal of the Royal Statistical Society: Series B (Methodological)* **53** 269–283.

Smith, T. and Vounatsou, P. (2003). Estimation of infection and recovery rates for highly polymorphic parasites when detectability is imperfect, using hidden Markov models. *Statistics in Medicine* **22** 1709–1724.

Spencer, S. E. F., Besser, T. E., Cobbold, R. N. and French, N. P. (2015). 'Super' or just 'above average'? Supershedders and the transmission of *Escherichia coli* O157:H7 among feedlot cattle. *Journal of the Royal Society Interface* **12** 0446.

Streftaris, G. and Gibson, G. J. (2004). Bayesian inference for stochastic epidemics in closed populations. *Statistical Modelling* **4** 63-75.

R Core Team (2019). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Tenover, F., Arbeit, R., Goering, R., Murray, B., Persing, D., Pfaller, M. and Weinstein, R. (1997). How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: A review for healthcare epidemiologists. *Infection Control and Hospital Epidemiology* **18** 426–439.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface* **6** 187–202.

Touloupou, P. (2016). Bayesian inference and model selection for partially observed stochastic epidemics, PhD thesis, University of Warwick.

Touloupou, P., Finkenstädt, B. and Spencer, S. E. F. (2019). Scalable Bayesian inference for coupled hidden Markov and semi-Markov models. *Journal of Computational and Graphical Statistics* 1–12.

Touloupou, P. and Spencer, S. E. F. (2020). epiPOMS: Bayesian Inference for Partially Observed Multi-Strain Epidemics R package, version 0.1.0.

Worby, C. J., O'Neill, P. D., Kypraios, T., Robotham, J. V., De Angelis, D., Cartwright, E. J., Peacock, S. J. and Cooper, B. S. (2016). Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *The Annals of Applied Statistics* **10** 395-417.

Xu, X., Kypraios, T. and O'Neill, P. D. (2016). Bayesian non-parametric inference for stochastic

epidemic models using Gaussian Processes. *Biostatistics* **17** 619–633.

P. Touloupou
B. Finkenstädt
S.E.F. Spencer
Department of Statistics,
Zeeman Institute for Systems Biology
and Infectious Disease Epidemiology Research
University of Warwick
Coventry, UK
E-mail: P.Touloupou.1@warwick.ac.uk
B.F.Finkenstadt@warwick.ac.uk
S.E.F.Spencer@warwick.ac.uk

T.E. Besser
Department Veterinary Microbiology
and Pathology
Washington State University
Pullman, WA 99164, USA
E-mail: tbesser@wsu.edu

N.P. French
The New Zealand Food Safety
Science and Research Centre
School of Veterinary Science
Massey University
Palmerston North, New Zealand
E-mail: N.P.French@massey.ac.nz