Mathematical Medicine and Biology (2018) **35** (Supplement 1), i29–i49
doi: 10.1093/imammb/dqx010
Advance Access publication on 12 October 2017

Extended models for nosocomial infection: parameter estimation and model selection

ALUN THOMAS*
Division of Genetic Epidemiology, School of Medicine, University of Utah, 391 Chipeta Way Suite D, Salt Lake City, UT 84108, USA
*Corresponding author. Email: alun@genepi.med.utah.edu

KARIM KHADER, ANDREW REDD, MOLLY LEECASTER, YUE ZHANG, MAKOTO JONES, TOM GREENE AND MATTHEW SAMORE
Division of Epidemiology, School of Medicine, University of Utah, 295 Chipeta Way, Salt Lake City, UT 84132, USA and VA Salt Lake City Health Care System, 500 Foothill Drive Building 182, Salt Lake City, UT, 84148, USA

[Received on 10 November 2015; revised on 18 May 2017; accepted on 18 August 2017]

We consider extensions to previous models for patient level nosocomial infection in several ways, provide a specification of the likelihoods for these new models, specify new update steps required for stochastic integration, and provide programs that implement these methods to obtain parameter estimates and model choice statistics. Previous susceptible-infected models are extended to allow for a latent period between initial exposure to the pathogen and the patient becoming themselves infectious, and the possibility of decolonization. We allow for multiple facilities, such as acute care hospitals or long-term care facilities and nursing homes, and for multiple units or wards within a facility. Patient transfers between units and facilities are tracked and accounted for in the models so that direct importation of a colonized individual from one facility or unit to another might be inferred. We allow for constant transmission rates, rates that depend on the number of colonized individuals in a unit or facility, or rates that depend on the proportion of colonized individuals. Statistical analysis is done in a Bayesian framework using Markov chain Monte Carlo methods to obtain a sample of parameter values from their joint posterior distribution. Cross validation, deviance information criterion and widely applicable information criterion approaches to model choice fit very naturally into this framework and we have implemented all three. We illustrate our methods by considering model selection issues and parameter estimation for data on methicilin-resistant *Staphylococcus aureus* surveillance tests over 1 year at a Veterans Administration hospital comprising seven wards.

Keywords: MRSA infection; patient level models; reversible jump MCMC; information criteria; cross validation.

1. Introduction

A wide variety of microorganisms are transmitted within healthcare settings, posing significant threats to both patients and healthcare workers. Some pathogens, such as SARS associated coronavirus and Ebola, have the potential to disseminate widely in community populations. Other types of emergent threats are represented by antibiotic-resistant bacteria, such as methicillin-resistant *Staphylococcus aureus* (*MRSA*) or vancomycin-resistant *Enterococcus*, many of which have become endemic in hospitals. Although a substantial body of knowledge has accumulated about the epidemiology of nosocomial, or hospital associated, infections, much remains unknown about modes of transmission. For instance, is acquisition of a pathogen a consequence of direct patient to patient contact, is it mediated by health care workers, or is it the result of general environmental contamination? Hospitals routinely collect and maintain extensive records of patient admission and discharge times to and from wards and units, and records are also available for the sample times and results of laboratory tests to detect pathogens colonizing and infecting patients. Colonization generally refers to the situation where bacteria are living on the skin or in other body sites without causing disease. Our usage of the term also refers to the state in which an individual is shedding an organism and capable of transmitting to others, i.e. is infectious. Clinical infection is disease caused by a microorganism. For bacteria that cause nosocomial infection, colonization typically precedes clinical infection. Because we do not analyse clinical data here, we are unable to distinguish asymptomatic colonization from clinical infection. The ultimate objective of our work is to reduce the rates of nosocomial infection. Hospital records have been used in a variety of ways to estimate rates of colonization and hence measure the effect of interventions designed to reduce acquisition rates. Our work focuses on patient level models as opposed to compartmental models (Cooper & Lipsitch, 2004; McBryde *et al.*, 2007), and is thus very much in the vein of Forrester *et al.* (2007) and Cooper *et al.* (2008), and more recently, Kypraios *et al.* (2010), Worby *et al.* (2013) and Haverkate *et al.* (2015).

Patient level nosocomial infection modelling is a classical hidden variable statistical problem: although swab tests and admission data are informative, the precise time of colonization is very unlikely to be observed. Cooper *et al.* (2008) briefly reviewed previous attempts to address the problem before introducing an augmented data approach. In this, the observed data are augmented by the unobservable times of colonization, the values of which are repeatedly imputed based on current parameter values and the observed data. The imputations and parameter updates are made using Markov chain Monte Carlo stochastic integration (*MCMC*) which, with care, ensures that the parameter values so generated constitute a correlated sample from the posterior densities of the parameters in a Bayesian analysis. These methods can typically also be adapted for stochastic optimization to find maximum likelihood or maximum posterior estimates, but we will not consider this further in this work.

The benefit of the hidden variable approach is that, given the augmented data values, the estimation problem is greatly simplified. The cost is that many samplings of the augmented data need to be made which may be computationally expensive and often intricate to program. Although the model of Cooper *et al.* (2008) was specified in continuous time, their program implemented a discrete time approximation. Thomas *et al.* (2015) gave a somewhat more efficient discrete time implementation, and provided a true continuous time solution. They showed that the continuous implementation was no more computationally demanding than the discrete time one, and moreover, eliminated some biases in parameter estimation. The programs we provide for the analyses described here are implemented in continuous time.

The novel contribution of this work is to extend the susceptible-infected, or *SI*, model which assumes that once colonized an individual immediately becomes infectious and remains so for the duration of the study, to models that allow for decolonization or loss of infectiousness, i.e. susceptible-infected-susceptible or *SIS* models, and for a latent period following exposure and preceding infectiousness, i.e. susceptible-exposed-infected-susceptible or *SEIS* models. The model assumption that the infectious state never resolves conflicts with the current body of knowledge about the nature of colonization. Substantial empirical evidence has accumulated that a variety of types of bacteria which colonize human hosts, including MRSA, can be lost and reacquired during follow-up (Huckabee *et al.*, 2009; Haverkate *et al.*, 2014; Shenoy *et al.*, 2014). Rather than assuming that every negative surveillance test that occurs after a positive test is always a false negative, it is much more realistic to allow the possibility of decolonization. For the most part, information for estimating the decolonization rate comes from tests done on patients with previous positives tests, hence, this will be better estimated under a regime of regular surveillance testing. Incorporating a latent period enhances the realism of the model as, in nature, there is always some lag between transmission and onset of infectiousness in the newly colonized host. For SIS and SEIS models the number of transitions in different imputations of the augmented data can vary, and, thus, so will the dimensionality of the problem. We address this using reversible jump MCMC, or *RJMCMC*, updates for the patient histories that constitute the augmented data (Green, 1995).

We now also allow for data from multiple facilities and multiple wards or units within a facility. The movements of patients between units and time spent outside of any unit are tracked and the colonization status at readmission will depend on the status at previous discharge and the time between these events. A two- or three-state continuous time Markov process is used to model the out of facility colonization process depending on whether or not the model allows for a latent state. Models that assume that no patients are hospitalized more than once fail to account for the reality that readmission occurs with appreciable frequency. The reason this simplifying assumption is problematic is that readmission provides the key link between past transmission and future importation. The notion that this is an epidemiologically important relationship has received support in previous studies of MRSA colonization (Jones *et al.*, 2015): lagged facility-level MRSA acquisition was found to be temporally correlated with admission prevalence. The out of unit decolonization rate is informed by the colonization status on readmission for patients likely to have been colonized on their discharge, and will decrease at the the time interval between hospitalization increases.

We also introduce a new parameter for the probability that a patient in the unit at the onset of the study is colonized. In previous models such individuals were typically treated as time zero admissions, and while the practical effect of this is likely to be minimal, it is clearly an inappropriate assumption as the patient might be colonized as the result of previous in unit exposure. We note that we regard this as a nuisance parameter introduced to avoid problems in estimating parameters of interest rather than something interesting in its own right.

Finally, we consider three types of patient-to-patient transmission models: models where the hazard to a patient of becoming colonized is constant; is proportional to the number of colonized patients in the unit and/or facility, usually called the *density dependent model*; or is proportional to the proportion of colonized patients, usually called the *frequency dependent model* (McCallum *et al.*, 2001).

In order to inform model choice, we follow Forrester *et al.* (2007) in using cross validation, Gelman *et al.* (2014) and Cooper *et al.* (2008) in using the deviance information criterion (*DIC*), (Spiegelhalter *et al.*, 2002) as both fit very naturally into our framework, although both require additional MCMC runs over and above those needed for parameter estimation. We make the slight extension from leave one out cross validation to a version that allows for arbitrary sets of tests to be omitted in the estimation phase before being predicted. Celeux *et al.* (2006) consider several possible versions of the DIC for missing data problems which treat the augmented variables variously as data or parameters. As the DIC requires calculation of deviances under fixed parameter estimates, we find it inappropriate in a context where the augmented variables are complete patient histories rather than simple numerical quantities, so that using a posterior mean estimate for example is not well defined. We, therefore, have chosen to implement what Celeux *et al.* (2006) refer to as DIC_6, a version that treats the augmented variables as data. We also consider the *widely applicable, or Watanabe–Akaike, information criteria (WAIC)* (Watanabe, 2010). This has the computational attraction of being calculable from statistics generated in the course of the simulation run for parameter estimation, thus, unlike cross validation and DIC it requires minimal additional effort. We will explore these model selection approaches in the example we have chosen to illustrate our methods. This will be an analysis of MRSA surveillance test results over a 1-year period from a Veterans Administration hospital comprising seven different units.

2. Methods

2.1 *Observable data and events*

The hospital record information required for estimating transmission model parameters can conveniently be specified as a list of events, each specified by five fields: unique integer identifiers for the facility, unit and patient, a decimal specifying the time of the event and an integer code giving the type of the event. As we track transfers of patients between facilities and units, it is essential that their identifiers are uniquely defined throughout the entire data set, not just unique within a unit. Times are expressed as decimals in days to arbitrary precision and relative to a common reference time. By facility we mean a hospital, long-term care facility, or similar and units within a facility may include, for instance, wards or intensive care units.

Observed events include admissions, discharges and tests. We identify patients present in a unit at the beginning of a study and indicate these with what we refer to as an in situ event and distinguish these from true admission events that occur subsequently. Test events are taken to occur at the time the test samples were obtained, not the time at which the results became available. On input, the event list is checked for consistency so that the admission and discharge times specify a coherent history of inpatient episodes, and that test events only occur during an episode. Multiple inpatient episodes for the same individual are identified and linked. If time data are not sufficiently precise, for instance, if reported only to the nearest day, then concurrent events are sorted so that admissions and discharges make a coherent history. Concurrent negative and positive test are sorted arbitrarily. We prepend and append events indicating the start and stop of the study to the list.

2.2 *Unobservable events and augmented data*

The unobservable elements that complete the augmented data set in this hidden Markov model comprise the state of a patient on admission to a unit, and subsequent changes in that state. In the three state SEIS model, a patient may be uncolonized, in the latent period, or colonized and infectious. Acquisition, progression and decolonization events, respectively, indicate the transitions out of these states and into the next in a cyclical pattern. Individuals in the latent state are assumed not to be infectious, and should be negative in any tests. In the two state SIS model, there is no latent period and individuals move directly from uncolonized to colonized on acquisition.

At the beginning of the MCMC simulation, these unobserved variables specifying each patient's underlying colonization status are initialized arbitrarily subject only to their being colonized at the time of any positive test, so that the initial configuration has positive probability under models that do not allow for false-positive tests. At subsequent MCMC updates, we use the Metropolis–Hastings method proposing a new colonization history for each individual inpatient episode and accepting or rejecting with the appropriate probability.

2.3 *Models and parameters*

We follow Thomas *et al.* (2015) in expressing the likelihood for a list of events D containing the augmented data given the model parameters θ, as product of event terms and gap terms

$$\pi(D|\theta) = \prod_{e \in D} g(e;\theta)h(e^-, e;\theta) \tag{1}$$

where the product is taken over all events other than the start event which has no predecessor. The function $h(e^-, e;\theta)$ is the probability that none of the potential events that could have occurred in the gap between e^-, the previous event, and e, the current event, did occur, and $g(e;\theta)$ is the probability or hazard of the event e as appropriate for the type of event.

The information required to calculate this likelihood will vary depending on the choice of model, however, for all the models considered below the following is sufficient. For each event e we have:

- $t(e)$: The time of the event.
- $s(e)$: For events with an associated patient, the status, uncolonized, latent or colonized, denoted $\{0, 1, 2\}$ respectively, of the patient immediately after the event.
- $\{n_{i,j,k}(e)\}$: The number of patients in facility i, unit j, who are in colonization state k immediately after the event. Let $n_{i,j}(e) = \sum_k n_{i,j,k}(e)$.

Much of the programming effort in implementing these analyses is in constructing data structures that capture this information, and efficiently maintain it in the course of MCMC updates.

The form of the functions $g()$ and $h()$ will also, of course, depend on the model. We will describe in detail these functions for a three-state model with unit specific colonization rates $\{\lambda_{i,j}\}$ following the density dependent form. We will assume that all progressions and decolonizations occur at random at common rates ρ and δ for each patient in the appropriate state, and that σ_k is the probability that a patient in situ at the beginning of the study is in colonization state k. We will further assume that false negative tests occur with common probability ϕ, that there are no false positives, and that the testing process is independent of the underlying colonization state of a patient. Transitions between colonization states for individuals not currently within the system are assumed to follow a three-state cyclical Markov process with transition rates κ, μ and ν out of the uncolonized, latent and colonized states, respectively. We will discuss only briefly the steps required for variations of this model that we have also implemented.

The gap terms can be calculated as

$$\log h(e^-, e) = [t(e) - t(e^-)] \sum_{i,j} n_{i,j,0}(e^-)n_{i,j,2}(e^-)\lambda_{i,j} + n_{i,j,1}(e^-)\rho + n_{i,j,2}(e^-)\delta. \tag{2}$$

The event specific term, $g(e;\theta)$, is calculated as follows for each type of event:

$$g(e;\theta) = n_{i,j,2}(e^-)\lambda_{i,j} \quad \text{if } e \text{ is a colonization} \tag{3}$$

$$\rho \quad \text{if } e \text{ is a progression} \tag{4}$$

$$\delta \quad \text{if } e \text{ is a decolonization} \tag{5}$$

$$\phi \quad \text{if } s(e) = 2 \text{ and } e \text{ is a negative test} \tag{6}$$

$$(1 - \phi) \quad \text{if } s(e) = 2 \text{ and } e \text{ is a positive test} \tag{7}$$

$$\sigma_{s(e)} \quad \text{if } e \text{ is an in situ event} \tag{8}$$

$$\tau(t(e) - t(e^d), s(e^d), s(e)) \quad \text{if } e \text{ is an admission} \tag{9}$$

$$1 \quad \text{otherwise.} \tag{10}$$

When e is a patient's admission, e^d is the discharge event for that patient's most recent previous inpatient episode, which may have been in the same facility or unit or in different ones. If this is the patient's first admission into the system, e^d will be null and dealt with as described below. The function $\tau(t, u, v)$ is the probability that a patient who is discharged from a unit in state u is in state v t days later when they are readmitted or admitted to another unit. These are obtained from the assumption that the out of unit process is the three-state cyclical Markov process with transition rate matrix

$$Q = \begin{pmatrix} -\kappa & \kappa & 0 \\ 0 & -\mu & \mu \\ \nu & 0 & -\nu \end{pmatrix} \tag{11}$$

and are given in matrix form by the matrix exponent e^{tQ}, which can be calculated using the Caley–Hamilton theorem as shown in the Appendix. For a patient's first admission to the system, we use the ergodic distribution given by

$$\frac{1}{\mu\nu + \nu\kappa + \kappa\mu} \begin{pmatrix} \mu\nu & \nu\kappa & \kappa\mu \end{pmatrix}. \tag{12}$$

This model was chosen as the simplest Markov process that reflects the same cyclical progression modeled for inpatients. While acquisition and decolonization rates for outpatients might well be different to those for inpatients, it would be reasonable to assume the same progression rate for both processes. In enforcing this constraint, however, we would lose the independence between inpatient and outpatient processes that we have if we condition on the augmented data, and, thus, complicate the parameter updating procedures. So for purely pragmatic reasons we do not enforce this constraint.

2.4 *Model variations*

For frequency dependent models, $\lambda_{i,j} n_{i,j,2}(e^-)$ is replaced with $\lambda_{i,j} n_{i,j,2}(e^-)/n_{i,j}(e^-)$ in both (2) and (3). For models with constant colonization it becomes simply $\lambda_{i,j}$.

For models that allow for transmission within facilities, as well as within units, additional terms and parameters are required for the gap probabilities, and (3) also changes.

It is straightforward to allow progression, decolonization and test parameters to depend on the facility or unit specific to the event, however, see the notes below on using such extensions. If the rate of testing is allowed to depend on the colonization status of the patient, this must also be accounted for in the gap term.

Using the two-state SIS model is straightforward: progression events are removed from the framework and patients transition immediately from uncolonized to colonized on acquisition. Equation (2) becomes

$$\log h(e^-, e) = [t(e) - t(e^-)] \sum_{i,j} n_{i,j,0}(e^-) n_{i,j,2}(e^-) \lambda_{i,j} + n_{i,j,2}(e^-)\delta. \tag{13}$$

The out of unit process simplifies greatly to give

$$\tau(t) = \frac{1}{\nu+\kappa}\begin{pmatrix}\nu & \kappa \\ \nu & \kappa\end{pmatrix} - \frac{e^{-t(\nu+\kappa)}}{\nu+\kappa}\begin{pmatrix}-\kappa & \kappa \\ \nu & -\nu\end{pmatrix} \tag{14}$$

with the first matrix giving the importation probabilities for the first entry into the system. This two-state process is different to that used by Cooper *et al.* (2008) who had the probability of being colonized at readmission going to zero as the time out of the system increased, whereas here it tends to $\frac{\nu}{\kappa+\nu}$. The SI model is implemented by fixing decolonization rates, in and out of unit, at zero.

2.5 *Parameter priors and MCMC updates*

The likelihood $\pi(D|\theta)$ and parameter prior $\pi(\theta) = \pi(\lambda, \rho, \delta, \phi, \sigma, \kappa, \mu, \nu)$ specify the parameter posterior distribution given the augmented data. We can sample from this posterior using MCMC updates for both the parameters given the augmented data, and the unobservable elements in the augmented data given the parameter values and observed data.

For the parameters that are probabilities we have implemented Beta or Dirichlet priors as these are conjugate with the appropriate elements of the likelihood making posterior sampling straightforward in most cases. For rate parameters, we assume Gamma priors as these are again in many cases conjugate with the likelihood and make posterior sampling with Gibbs updates easy.

Given the augmented data, the conditional distribution of several parameters are independent of the other parameters, and because of conjugacy, updates can be generated by Gibbs sampling (Geman & Geman, 1984) using simple counts of sufficient statistics. For instance, the in situ colonized probability σ and the false-negative probability ϕ can be updated in this way. In other cases, Metropolis's method is employed (Metropolis *et al.*, 1953) using symmetrical proposals following logit and log transformations, respectively, for probability and rate parameters. For any given state of the augmented data, the calculations required for Metropolis updates are very quick, so, to improve mixing we make 100, of these updates after each augmented data update. The parameters of the out of unit colonization process, $\{\kappa, \mu, \nu\}$, are always updated using Metropolis, as the form of the likelihood does not allow conjugacy.

2.6 *RJMCMC updates for the augmented data*

As with Cooper *et al.* (2008) and Thomas *et al.* (2015), we use stochastic integration to account for the uncertainty in the augmented data set by making a round of updates that sample from the conditional distribution of each patient's underlying colonization states given the observed data, the current parameter values, and the current history of all other patients. Thomas *et al.* (2015) parametrized their model in such a way that the dimensionality of the augmented data was constant. Cooper *et al.* (2008) framed their model so that the dimensionality changed depending on whether a patient experienced a colonization or not and used RJMCMC (Green, 1995) to account for this. We note in passing that since their model was in discrete time, RJMCMC was not strictly speaking necessary. As our model will allow for zero or multiple colonization, progression and decolonization events in a patient episode, and is a continuous time model, the dimensionality is truly variable and RJMCMC, which we employ, or an alternative such as birth-death MCMC (Stephens, 2000), is necessary.

The complete history for a single-patient episode can be specified by the initial state x, the number of times a change of state occurs y, and the sorted list of change times $u = (u_1, u_2, \dots u_y)$. Letting D^- represent the current state of the augmented data for all but the episode being updated, we can write the conditional posterior for this episode as $\pi(x, y, u|D^-, \theta)$ which can be calculated efficiently from the stored list of events. To implement an RJMCMC update, we first sample new values x' and y' such that $P(X = x) = \gamma_x$ for $x = 0, 1, 2$ and $Y \sim \text{Poisson}(\alpha)$. Atypically for RJMCMC, this proposal does not depend on the current state. The values of γ and α can be chosen arbitrarily to tune the mixing performance of the sampler.

We then sample y' sorted $U(t(a), t(d))$ random variables $v_1, \dots v_{y'}$, where a and d are the admission and discharge events for the episode. That is, we sample $(v_1, \dots v_{y'})$ from the probability density function

$$\frac{y'!}{(t(d) - t(a))^{y'}} \quad t(a) < v_1 < v_2 < \cdots < v_{y'} < t(d). \tag{15}$$

Thus, $(u_1, \dots u_y, v_1 \dots v_{y'})$ forms the complete set of continuous variables, active and auxiliary, associated with the current state. We transform this to the variables $(u'_1, \dots u'_{y'}, v'_1 \dots v'_y)$ required for the new state by setting $u'_i = v_i, i = 1 \dots y'$ and $v'_j = u_j, j = 1 \dots y$, in effect simply switching the role of the active and auxiliary variables. This transformation is clearly reversible and has Jacobian 1.

Hence, the proposed state x', y', u' is accepted with probability

$$\min\left\{1, \frac{\pi(x', y', u'|D^-, \theta)}{\pi(x, y, u|D^-, \theta)} \left(\frac{\gamma_x}{\gamma_{x'}}\right) \left(\frac{\alpha}{t(d) - t(a)}\right)^{y-y'}\right\} \tag{16}$$

To calculate the ratio of the probabilities of the incumbent and proposed states we note that when $\pi(x', y', u'|D^-, \theta)$ and $\pi(x, y, u|D^-, \theta)$ are expanded using (1), with one exception, the terms for events occurring outside the interval $(t(a), t(d))$ cancel. The exception to this occurs when there is a subsequent inpatient episode for the patient whose episode is being updated. In this case the contribution from the admission event for that episode also has to be adjusted. Thus, computation of this posterior density is relatively efficient depending only the number of events that occur during the episode, and is independent of the length of the study.

2.7 *Model selection methods*

We have implemented, and will compare below, three methods for model selection: cross validation, the DIC and the WAIC. Here we let y denote the observed test result data, and X, the hidden variables specifying the underlying patient colonization states.

To evaluate the cross validation predictive probability, we consider leaving out sets of test results indexed by I, running the sampler using only y_{-I} the remaining test data and predicting each omitted test result y_i from the simulated augmented data and parameters. That is, we evaluate the mean cross validation predictive probability

$$\begin{aligned} \text{CVPP} &= \frac{1}{n} \sum_I \sum_{i \in I} \pi(y_i|y_{-I}) && (17) \\ &= \frac{1}{n} \sum_I \sum_{i \in I} \int_{X,\theta} \pi(y_i|X, \theta)\pi(X, \theta|y_{-I}) dX d\theta && (18) \\ &\approx \frac{1}{n} \sum_I \sum_{i \in I} \frac{1}{S} \sum_{s=1}^{S} \pi(y_i|X^s, \theta^s) && (19) \end{aligned}$$

where $X^s, \theta^s \sim \pi(X, \theta | y_{-I})$. In our analysis below we report LCVPP $= -log$(CVPP) so that, like the DIC and WAIC, we prefer small values.

In its standard form the DIC is given by

$$DIC = -2 \log \pi(y|\bar{\theta}) + 2p_d \tag{20}$$

where p_d is a measure of the effective number of parameters

$$p_d = -2 \int_\theta \log [\pi(y|\theta)] \pi(\theta|y) d\theta + 2 \log \pi(y|\bar{\theta}) \tag{21}$$

and $\bar{\theta}$ is a set of parameter estimates, typically posterior means in a Bayesian framework. Since the augmented data are patient colonization histories rather than simple numerical quantities, their mean is not well defined, and hence neither would $-2 \log (\pi(y|\bar{X}, \bar{\theta}))$ the value of the deviance at their mean. Thus, we are not simply able to consider the augmented data as parameters in evaluating the DIC. Celeux *et al.* (2006) considered several variants of the DIC suitable for use with hidden variables. Of these, their DIC_6 defined as the mean of the DIC over the values of the hidden variables considered as data works most easily in our situation, and is the one we have chosen:

$$DIC = -2 \int_X \log [\pi(y, X|\bar{\theta})] \pi(X|y) dX + 2p_d \tag{22}$$

$$p_d = -2 \int_X \int_\theta \log [\pi(y, X|\theta)] \pi(\theta|y) \pi(X|\theta, y) dX d\theta + 2 \int_X \log [\pi(y, X|\bar{\theta})] \pi(X|y) dX \tag{23}$$

$$= -2 \int_{X,\theta} \log [\pi(y, X|\theta)] \pi(X, \theta|y) dX d\theta + 2 \int_X \log [\pi(y, X|\bar{\theta})] \pi(X|y) dX \tag{24}$$

which can be evaluated by simulation as

$$\int_{X,\theta} \log[\pi(y, X|\theta)] \pi(X, \theta|y) \approx \frac{1}{S} \sum_{s=1}^{S} \log \pi(y, X^s|\theta^s) \tag{25}$$

$$\int_X \log [\pi(y, X|\bar{\theta})] \pi(X|y) \approx \frac{1}{S} \sum_{s=1}^{S} \log \pi(y, X^s|\bar{\theta}) \tag{26}$$

with $X^s, \theta^s \sim \pi(X, \theta|y)$.

The WAIC is based on the log posterior predictive probability and, again, a penalty term p_W for the effective number of parameters

$$\text{WAIC} = -2 \sum_i \log \int_{X,\theta} \pi(y_i|X, \theta) \pi(X, \theta|y) dX d\theta + 2p_W \tag{27}$$

with p_W being obtained as

$$p_w = 2 \sum_i \log \int_{X,\theta} \pi(y_i|X, \theta) \pi(X, \theta|y) dX d\theta - 2 \sum_i \int_{X,\theta} \log [\pi(y_i|X, \theta)] \pi(X, \theta|y) dX d\theta. \tag{28}$$

or from the variance of the log posterior predictive probabilities

$$p_w = \frac{1}{4}\sum_i V[-2\log\pi(y_i|X,\theta)] \tag{29}$$

$$= \sum_i \left\{ \int_{X,\theta} (\log\pi(y_i|X,\theta))^2 \pi(X,\theta)dXd\theta - \left[\int_{X,\theta} \log\left(\pi(y_i|X,\theta)\right)\pi(X,\theta)dXd\theta \right]^2 \right\}. \tag{30}$$

We evaluate these by simulation using

$$\int_{X,\theta} \pi(y_i|X,\theta)\pi(X,\theta)dXd\theta \approx \frac{1}{S}\sum_{s=1}^{s} \pi(y_i|X^s,\theta^s) \tag{31}$$

$$\int_{X,\theta} \log\left[\pi(y_i|X,\theta)\right]\pi(X,\theta)dXd\theta \approx \frac{1}{S}\sum_{s=1}^{s} \log\pi(y_i|X^s,\theta^s) \tag{32}$$

$$\int_{X,\theta} [\log\pi(y_i|X,\theta)]^2\pi(X,\theta)dXd\theta \approx \frac{1}{S}\sum_{s=1}^{s} [\log\pi(y_i|X^s,\theta^s)]^2 \tag{33}$$

with again $X^s, \theta^s \sim \pi(X,\theta|y)$ in each case.

2.8 *Computational complexity*

The computational structure of the MCMC updates made here is the same as that done by Thomas *et al.* (2015), so, again, the computational time required is dominated by that needed to make a scan of updates to the patient histories. When a patient history is updated it affects the states associated with all events that take place during their inpatient stay which is roughly proportional to the number of other patients in the unit at that time. Thus, in the worst case, which is when each unit is at capacity, a complete sweep of updates to the patient histories and parameters can be done in time of $O(NC)$ where N is the number of patient episodes and C is the capacity of the unit. The storage requirement is simply the list of all events which is of length $O(N)$. The methods have been implemented in C++ programs that are available from the corresponding author.

For cross validation, since running the sampler to re-estimate X and θ for each set I is computationally demanding, we avoid leave one out cross validation and instead make 10 runs leaving out roughly 10% of the test results each time with each test result randomly allocated to one decile. We typically make an initial burn-in run of MCMC updates using all the test data to obtain initial parameter estimates. Then each of the 10 deciles is omitted, another burn-in run is made to allow the parameter estimates and augmented data to adjust, and then a further run is made under which the predictive probabilities are estimated. These latter 2 steps are repeated for each decile.

The DIC needs two MCMC runs for evaluation. The first to obtain parameter estimates $\bar{\theta}$. The second to obtain the likelihoods of the simulated complete data under these parameters and their likelihoods under the simulated parameter values.

For the WAIC, the augmented data are effectively considered as parameters and the integrals required, using either version of p_w, can be evaluated from the same run of MCMC simulations used to obtain parameter posterior distributions, making this the most computationally attractive option.

2.9 *Identifiability and over parametrization*

Our framework and programs enable a wide range of model extensions and are flexible enough to accommodate unit specific, facility specific or system wide parameters. Full exploitation of this, however, is of questionable value as it can lead to serious over parametrization and the possibility of parameters becoming unidentifiable. Even in situations where, given the augmented data, parameters can be independently estimated, there may be identifiability issues when we integrate over the hidden variables. Evidence of this was shown by Thomas *et al.* (2015) where negative correlation was seen between, for instance, estimates of importation probabilities and in unit colonization rates with different imputations of the hidden variables explaining the observed positive tests as due to either importation or colonization. While such issues could be addressed by constraints imposed by strongly informative prior distributions, we would advise against these extensions other than where strongly justified by the data. One such situation is when qualitatively different tests are used for pathogen detection. For instance, MRSA surveillance tests usually use nasal or dermal swabs while clinical tests may use blood.

3. Data analysis

We present an analysis of data including admission, discharge and transfer times for patients at a Veterans Affairs (VA) acute care hospital. Transfers are handled in our analysis as a discharge and concurrent admission to another unit. The data also include the times and results of MRSA surveillance tests which were carried out on patients at admission, discharge and transfer, and cover all patients that were admitted during a 1 year period beginning 1 January 2008. All data management and analysis was carried out within the VA Informatics and Computing Infrastructure environment. The acute care facility was made up of seven units which we order in decreasing total number of admissions: General Acute Medicine, Telemetry, General Surgery, Medical ICU, Surgical ICU, Intermediate Medicine and Neurology. Table 1 gives a summary of patient movements into and between the seven units, and also the number of positive and negative tests for each unit. The sequences of test results for each patient, up to their fifth test is given in Table 2. Figure 1 gives the daily counts of patients in each unit throughout the study period and also the proportion of current inpatients who had positive tests prior to that time.

In order to investigate appropriate modelling of this system, and potential differences in colonization rates by unit, we fitted the parameters for 18 different models. We considered the SI, SIS and SEIS models each with either a constant, a density dependent or a frequency dependent colonization rate. For each of these combinations, we considered models with a common colonization rate for all units and with unit specific colonization rates.

The MRSA tests were all nasal swabs surveillance tests and we have assumed the same false negative probability for all tests throughout the facility. Similarly, we have assumed a common out of unit Markov process for all units.

All binomial and multinomial parameters were assigned uniform or uniform Dirichlet priors, and all rate parameters were assigned exponential unit rate priors.

In order to calculate parameter estimates and the WAIC, an MCMC run of 5000 iterations was made for each model. The first 1000 iterations were discarded and a sample of parameter values was collected from the remaining 4000. The parameters were estimated using the means of the sample. The posterior predictive probabilities needed to calculate either version of the WAIC were also estimated from these 4000 simulations. We then made a further run of 4000 simulations that were used to compute the differences between deviances under the simulated parameter values and the overall mean parameter values as required for the DIC. DIC and WAIC results are presented in Table 3. Two versions of the WAIC

TABLE 1 *Summary of unit admission and transfer histories*

Unit		1	2	3	4	5	6	7	Total
Patients in situ		16	12	7	7	4	2	0	48
New admissions		875	945	784	276	279	1	61	3221
Readmissions from unit									
Name	Number								
General Acute Medicine	1	274	178	132	136	62	93	8	883
Telemetry	2	304	193	74	102	60	23	9	765
General Surgery	3	137	81	117	38	108	145	1	627
Medical ICU	4	255	154	64	13	41	0	3	530
Surgical ICU	5	41	56	330	47	14	0	4	492
Intermediate Medicine	6	55	27	32	16	5	5	2	142
Neurology	7	6	6	7	2	9	10	6	46
Total admissions		1963	1652	1547	637	582	279	94	6754
Mean length of stay in days		3.32	2.46	2.80	3.47	3.39	4.09	3.03	3.04
Positive MRSA tests		369	173	158	93	46	33	10	882
Negative MRSA tests		3024	2525	2254	661	570	181	134	9349
Positives per patient day		0.057	0.043	0.036	0.042	0.023	0.029	0.035	0.043

are given: WAIC 1 uses (28) and WAIC 2 uses (30) to calculate p_w. Table 4 gives the estimates for the test false negative probability, the in situ status probabilities, and the state probabilities for the limiting case of the out of unit process for each model. Table 5 gives the estimates for the in unit colonization process rates.

For each model, a separate series of MCMC samples was generated in order to calculate the cross validation score. In each case, an initial run of 500 simulations was made with all test data included. One-tenth of the data was then omitted and 1000 more updates made. The first 500 were discarded and the second 500 were used to obtain predictive probabilities for the omitted results. This was repeated for each of the remaining 9 tenths. The log of the mean predictive probability is given in Table 3.

4. Discussion

Table 3 shows strong agreement between the LCVPP and the WAIC statistics. They both generally favour SEIS over SIS over SI models with the mean ranks for these classes of models over variants being 4.1, 8.9 and 15.5 respectively using LCVPP and 5.3, 7.6 and 15.5 using WAIC. Overall, the rank correlation between the two is 76%. Reassuringly, the 2 WAIC statistics give the models identical rankings, and correlate almost perfectly, so at least in this context, either is equally applicable. In contrast to the LCVPP and WAIC, the DIC favours SIS over SEIS over SI models with mean ranks 8.0, 5.8 and 14.7. The rank correlation between it and the LCVPP is only 40% and with the WAIC is 42%.

That the constant rate models do so well is unexpected. The LCVPP ranks the SEIS unit specific constant colonization rate as the best overall and ranks it best of the 3 models in each category. The

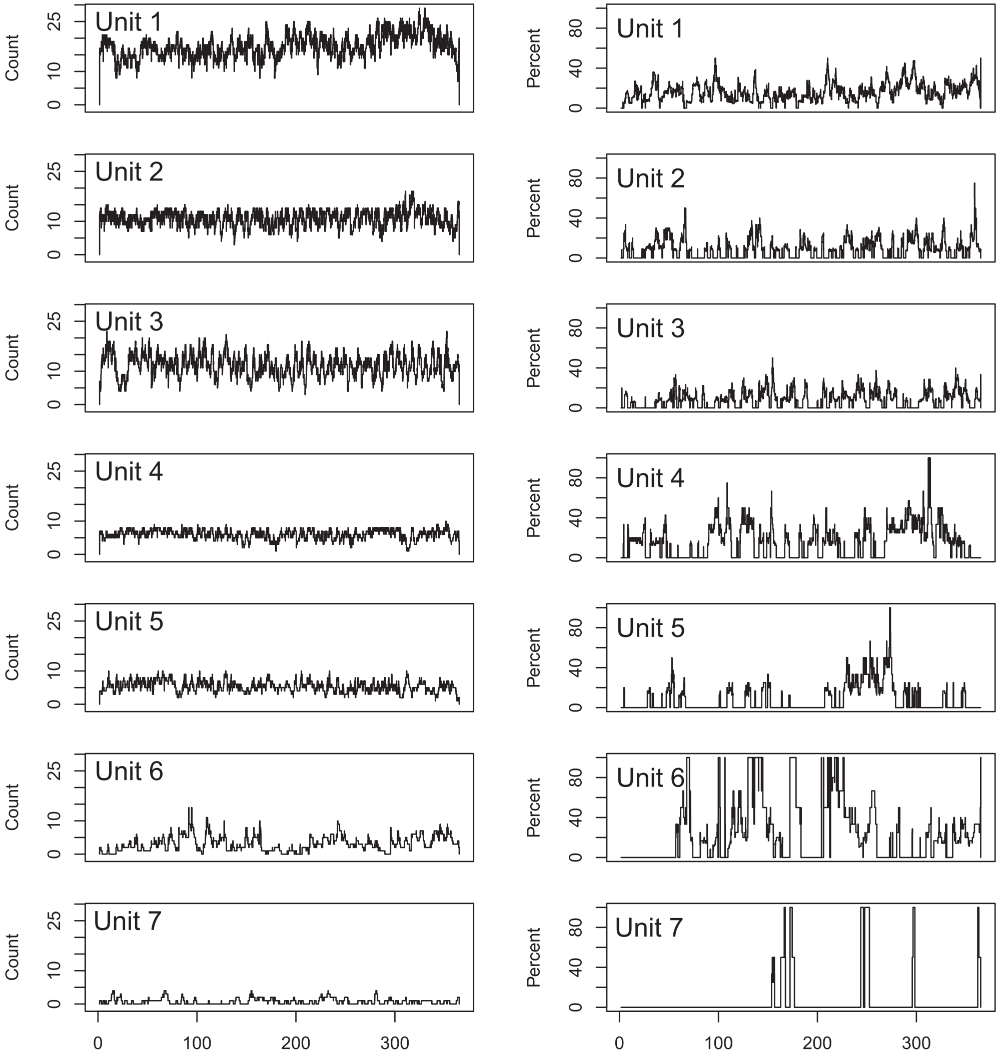

FIG. 1. The left column gives plots of each unit's inpatient count by day. The right column gives the percent of patients in the unit who had had a positive test during the study period prior to that day.

WAIC ranks the SEIS facility wide constant colonization rate model as the best overall. This brings into question somewhat the thinking behind the density and frequency dependent colonization assumptions. These are intended to model colonization by direct contact between colonized and susceptible patients when the number of potential interactions either grows linearly with the number of patients in the unit or is constant. Transmission of bacteria between patients is believed to be mediated primarily by contact involving healthcare workers and environmental surfaces rather than direct contact between patients, indeed, in many settings, particularly ICUs, patients are not mobile, so direct patient contact is unlikely. Thus, the environment can serve as a reservoir for acquisition after discharge or transfer of colonized

TABLE 2 *Summary of test result sequences. The sequences are for whole patient histories regardless of the number of inpatient episodes and the times between them. For five or more tests only the results of the first 5 are reported*

No tests

	39

One test

-	326
+	31

Two tests

−−	1400
−+	20
+−	9
++	80

Three tests

−−−	386
−−+	6
−+−	3
−++	6
+−−	7
+−+	3
++−	3
+++	32

Four tests

−−−−	307
−−−+	4
−−+−	1
−−++	5
−+−−	2
−+−+	1
−++−	0
−+++	5
+−−−	4
+−−+	1
+−+−	2
+−++	3
++−−	2
++−+	0
+++−	2
++++	14

Five or more tests

−−−−−	476
−−−−+	9
−−−+−	7
−−−++	9
−−+−−	4
−−+−+	0
−−++−	2
−−+++	4
−+−−−	2
−+−−+	0
−+−+−	1
−+−++	0
−++−−	0
−++−+	0
−+++−	1
−++++	3
+−−−−	3
+−−−+	0
+−−+−	0
+−−++	0
+−+−−	0
+−+−+	0
+−++−	0
+−+++	2
++−−−	6
++−−+	0
++−+−	2
++−++	3
+++−−	3
+++−+	2
++++−	8
+++++	18

patients. This likely depends on the cumulative recent history of colonized individuals in the unit, not simply on the instantaneous total. In modelling terms, it seems that there should be some lag in the system suggesting that moving average or other time series approaches should be investigated. A further factor may be that the mass action principle is less demonstrable if the prevalence or density of carriage shows little variation over time, and that distinguishing between frequency dependent and density dependent models will be difficult unless there is substantial variation in occupancy. Figure 1 shows that for the larger units, while there is substantial high frequency variation, occupancy numbers are generally stable, as are prevalence of carriage. Units 5, 6 and 7 show more variability, and some evidence of discrete

TABLE 3 *Model choice statistics*

Model			LCVPP		DIC		WAIC 1		WAIC 2	
			Value	Rank	Value	Rank	Value	Rank	Value	Rank
Two states. SI	Single colonization rate	Cons	0.09048	13	267080	10	0.26611	13	0.28631	13
		Dens	0.10830	18	293194	15	0.28687	15	0.30653	15
		Freq	0.09966	17	302150	17	0.29778	17	0.31758	17
	Unit specific colonization rates	Cons	0.09095	14	273276	12	0.27298	14	0.29292	14
		Dens	0.09859	15	322853	18	0.31917	18	0.34012	18
		Freq	0.09933	16	294968	16	0.28873	16	0.30839	16
Two states. SIS	Single colonization rate	Cons	0.08666	10	264882	2	0.23875	8	0.26129	8
		Dens	0.08665	8.5	265760	7	0.23617	3	0.25906	3
		Freq	0.08694	11	265547	6	0.24002	11	0.26243	11
	Unit specific colonization rates	Cons	0.08561	5	266260	8	0.23929	10	0.26176	10
		Dens	0.08641	7	266456	9	0.23696	5	0.25973	5
		Freq	0.08714	12	265073	3	0.23916	9	0.26168	9
Three states. SEIS	Single colonization rate	Cons	0.08418	4	274359	13	0.23538	1	0.25835	1
		Dens	0.08630	6	265536	5	0.23663	4	0.25945	4
		Freq	0.08665	8.5	263753	1	0.23755	7	0.26025	7
	Unit specific colonization rates	Cons	0.08206	1	272803	11	0.23723	6	0.25990	6
		Dens	0.08390	2	265144	4	0.24148	12	0.26371	12
		Freq	0.08416	3	274415	14	0.23576	2	0.25864	2

outbreaks, however, the low numbers here limit statistical power. We conclude that because of the rarity of transmissions, distinguishing decisively between modes of transmission will require a larger study.

The LCVPP shows clear support for unit specific colonization rates, the top 3 models all being unit specific SEIS. The LCVPP also generally favours unit specific rates within SI and SIS categories. This is supported by a broad range of rate estimates across the units, as shown in Table 5. Neither the DIC nor the WAIC, however, follow the same pattern. We also note that rate estimates correlate strongly and negatively with the size of the unit. There are far fewer observations on which to base rate estimates in the smaller units, and hence these have far larger posterior standard deviations. As these distributions are asymmetric the larger variability seems to also pull up the posterior mean. This suggests that using log normal priors, where the mean and variance can change independently, may be more appropriate although this, being non-conjugate, complicates computation. Cooper *et al.* (2008) used the log normal as did Khader *et al.* (2017) in an application of hierarchical modelling to compare colonization rates in a multi-unit intervention trial. We also note that unit 7 draws most of its patients from outside the system, thus, any positive tests observed there cannot be attributed to transfers from other units. Again, given the scarcity of transmissions, larger samples will be needed to overcome the influence of the priors and better evaluate evidence for different colonization rates in the units.

The parameter estimates given in Table 4 are remarkably stable and show no evidence of identifiability issues. The largest effect is that the false-negative test probability is larger for the SI models. Since, under these models, any negative test following a positive test must be a false negative, this is not surprising.

TABLE 4 *Test, in situ, and out of unit process parameter estimates. The estimates are the posterior means of a sample of 4000 observations. Estimates of the standard deviation of the posterior parameter marginals are given in italic text*

			Test false	In situ			Out of unit limiting		
			negative	state probabilities			state probabilities		
	Model		probability	0	1	2	0	1	2
	Single	Cons	0.443	0.725	—	0.275	0.871	—	0.129
			0.014	*0.071*	—	*0.071*	*0.006*	—	*0.006*
	colonization	Dens	0.480	0.712	—	0.288	0.849	—	0.151
			0.013	*0.073*	—	*0.073*	*0.007*	—	*0.007*
Two	rate	Freq	0.500	0.710	—	0.290	0.842	—	0.158
			0.013	*0.074*	—	*0.074*	*0.007*	—	*0.007*
states.	Unit specific	Cons	0.455	0.718	—	0.282	0.867	—	0.133
			0.013	*0.071*	—	*0.071*	*0.007*	—	*0.007*
SI	colonization	Dens	0.534	0.678	—	0.322	0.829	—	0.171
			0.013	*0.077*	—	*0.077*	*0.008*	—	*0.008*
	rates	Freq	0.484	0.709	—	0.291	0.847	—	0.153
			0.013	*0.072*	—	*0.072*	*0.007*	—	*0.007*
	Single	Cons	0.390	0.754	—	0.246	0.880	—	0.120
			0.015	*0.072*	—	*0.072*	*0.006*	—	*0.006*
	colonization	Dens	0.384	0.759	—	0.241	0.881	—	0.119
			0.015	*0.075*	—	*0.075*	*0.006*	—	*0.006*
Two	rate	Freq	0.392	0.754	—	0.246	0.880	—	0.120
			0.015	*0.075*	—	*0.075*	*0.006*	—	*0.006*
states.	Unit specific	Cons	0.391	0.769	—	0.231	0.880	—	0.120
			0.015	*0.073*	—	*0.073*	*0.006*	—	*0.006*
SIS	colonization	Dens	0.386	0.756	—	0.244	0.881	—	0.119
			0.015	*0.073*	—	*0.073*	*0.006*	—	*0.006*
	rates	Freq	0.390	0.759	—	0.241	0.881	—	0.119
			0.015	*0.072*	—	*0.072*	*0.007*	—	*0.007*
	Single	Cons	0.383	0.754	0.062	0.184	0.879	0.003	0.118
			0.015	*0.072*	*0.055*	*0.080*	*0.006*	*0.001*	*0.006*
	colonization	Dens	0.385	0.753	0.035	0.212	0.880	0.001	0.119
			0.015	*0.070*	*0.032*	*0.066*	*0.006*	*0.001*	*0.006*
Three	rate	Freq	0.387	0.754	0.048	0.199	0.881	0.000	0.119
			0.015	*0.071*	*0.038*	*0.068*	*0.006*	*0.000*	*0.006*
states.	Unit specific	Cons	0.387	0.760	0.051	0.190	0.880	0.001	0.119
			0.015	*0.066*	*0.048*	*0.069*	*0.006*	*0.001*	*0.006*
SEIS	colonization	Dens	0.395	0.709	0.079	0.212	0.880	0.001	0.119
			0.015	*0.074*	*0.053*	*0.070*	*0.006*	*0.001*	*0.006*
	rates	Freq	0.384	0.753	0.048	0.199	0.880	0.002	0.118
			0.015	*0.073*	*0.043*	*0.068*	*0.006*	*0.001*	*0.006*

Table 5 *In unit colonization process parameter estimates. The estimates are the posterior means of a sample of 4000 observations. Estimates of the standard deviation of the posterior parameter marginals are given in italic text. All rates and standard deviations have been multiplied by 1000000*

			Prog-	Decolo-		Colonization rate by unit						
	Model		ression	nization	All	1	2	3	4	5	6	7
	Single	Cons	—	—	60	—	—	—	—	—	—	—
			—	—	*62*	—	—	—	—	—	—	—
	colonization	Dens	—	—	29	—	—	—	—	—	—	—
			—	—	*29*	—	—	—	—	—	—	—
Two	rate	Freq	—	—	348	—	—	—	—	—	—	—
			—	—	*353*	—	—	—	—	—	—	—
states.	Unit specific	Cons	—	—	—	30	41	33	78	81	159	553
			—	—	—	*78*	*112*	*88*	*216*	*226*	*422*	*1413*
SI	colonization	Dens	—	—	—	7	19	25	48	92	169	3619
			—	—	—	*18*	*55*	*62*	*141*	*265*	*489*	*9731*
	rates	Freq	—	—	—	132	305	305	443	735	1008	9129
			—	—	—	*366*	*806*	*775*	*1206*	*1709*	*2482*	*23717*
	Single	Cons	—	1391	59	—	—	—	—	—	—	—
			—	*826*	*60*	—	—	—	—	—	—	—
	colonization	Dens	—	1670	30	—	—	—	—	—	—	—
			—	*1071*	*30*	—	—	—	—	—	—	—
Two	rate	Freq	—	1028	379	—	—	—	—	—	—	—
			—	*942*	*376*	—	—	—	—	—	—	—
states.	Unit specific	Cons	—	1346	—	28	40	39	93	97	184	557
			—	*1006*	—	*69*	*104*	*110*	*229*	*241*	*476*	*1452*
SIS	colonization	Dens	—	923	—	7	31	33	63	118	174	8319
			—	*877*	—	*19*	*75*	*77*	*170*	*327*	*446*	*20066*
	rates	Freq	—	962	—	138	366	381	417	806	758	17374
			—	*855*	—	*368*	*920*	*979*	*1096*	*2020*	*1932*	*47784*
	Single	Cons	719497	1614	56	—	—	—	—	—	—	—
			302153	*942*	*56*	—	—	—	—	—	—	—
	colonization	Dens	317847	1119	30	—	—	—	—	—	—	—
			218257	*812*	*30*	—	—	—	—	—	—	—
Three	rate	Freq	873119	594	390	—	—	—	—	—	—	—
			790277	*643*	*392*	—	—	—	—	—	—	—
states.	Unit specific	Cons	1107499	792	—	29	42	37	82	78	181	679
			695752	*699*	—	*72*	*110*	*93*	*234*	*225*	*486*	*1645*
SEIS	colonization	Dens	466430	565	—	7	29	26	64	138	159	7478
			454764	*565*	—	*20*	*75*	*70*	*164*	*377*	*423*	*20274*
	rates	Freq	487941	495	—	150	426	315	432	740	778	18408
			203982	*512*	—	*371*	*1066*	*829*	*1316*	*2023*	*2174*	*49470*

While we have not made a formal analysis of Table 2, and note that for brevity we have curtailed the sequence runs at 5, informal inspection reveals more sequences that follow the *positive–negative* pattern, favouring decolonization, than follow the *positive–negative–positive* pattern that would more strongly favour a higher false-negative rate. The average time spent in the latent state in SEIS models ranges from around 0.9 days for the constant model to around 3.2 days for the density dependent model. Estimates for the mean latent period for the outpatient process were broadly in line but showed more variability ranging from 0.8 to 13.1 days. Given that no data is collected during outpatient periods, the increased variability is not surprising. Although we consider the latency between colonization and detectability or infectiousness, as opposed to the more usual latency between acquisition and clinical infection, we would expect these to be of comparable magnitude and our estimates are consistent with what is known about MRSA (Dancer *et al.*, 2006). Nonetheless, we hesitate to interpret these estimates too directly. It may be that any improvement in predictive power of the SEIS model over the SIS is less to do with the existence of a latent state per se, and more to do with the slight lag it introduces into the system as discussed above.

The correlation of the LCVPP with the WAIC is far better than with the DIC, and given that the WAIC can be calculated at the same time as the parameter posterior samples are generated, we see no reason to pursue the DIC any further in this context. However, there is enough disagreement between LCVPP and WAIC that, despite theoretical results of asymptotic equivalence and ease of computation, the additional effort required for cross validation by decimation is indeed worthwhile, and the LCVPP statistic remains the best basis for model selection. This may be a situation where fitted and predictive errors converge slowly due to the discrete nature of our test result data having a strong effect on the imputation of the augmented data.

The programs used in this work are available from the corresponding author. While they are written to handle multiple facilities in a general way, adapting them for any specific sets of models would require some C++ programming. We hope in future to provide a more accessible method of using them, perhaps using the R statistical environment (R Core Team, 2015).

Acknowledgements

A.T. would like to thank Peter Green for time spent in Bristol and Sydney discussing RJMCMC and graphical modelling.

Funding

This investigation was supported by the University of Utah Study Design and Biostatistics Center, with funding in part from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant 5UL1TR001067-02 (formerly 8UL1TR000105 and UL1RR025764). This work was supported by COIN grant CIN 13-414 and CDC Epicenter grant 5U01CI000334.

References

Celeux, G., Forbes, F., Robert, C. P. & Titterington, D. M. (2006) Deviance information criteria for missing data models. *Bayesian Anal.,* **1**, 641–664.

Cooper, B. & Lipsitch, M. (2004) The analysis of hospital infection using hidden Markov models. *Biostatistics,* **5**, 223–237.

Cooper, B. S., Medley, G. F., Bradley, S. J. & Scott, G. M. (2008) An augmented data method for the analysis of nosocomial infection data. *Am. J. Epidemiol.,* **168**, 548–557.

Dancer, S. J., Coyne, M., Speekenbrink, A., Samavedam, S., Kennedy, J. & Wallace, P. G. M. (2006) MRSA acquisition in an intensive care unit. *Am. J. Infect. Control.,* **34**, 10–17.

Forrester, M. L., Pettitt, A. N. & Gibson, G. J. (2007) Bayesian inference of hospital-acquired infectious diseases and control measures given imperfect surveillance data. *Biostatistics* **8**, 383–401.

Gelman, A., Hwang, J. & Vehtari, A. (2014) Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24**, 997–1016.

Geman, S. & Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.,* **45**, 721–741.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Haverkate, M. R., Bootsma, M. C., Weiner, S., Blom, D., Lin, M. Y., Lolans, K., Moore, N. M., Lyles, R. D., Weinstein, R. A., Bonten, M. J. & Hayden, M. K. (2015) Modeling spread of KPC-producing bacteria in long term acute care hospitals in the Chicago region, USA. *Infect. Control. Hosp. Epidemiol.*, **36**, 1148–1154.

Haverkate, M. R., Derde, L. P. G., Brun-Buisson, C., Bonten, M. J. M. & Bootsma, M. C. J. (2014) Duration of colonization with antimicrobial-resistant bacteria after ICU discharge. *Intensive Care Med.,* **40**, 564–71.

Huckabee, C. M., Huskins, W. & Murray, P. R. (2009) Predicting clearance of colonization with vancomycin-resistant enterococci and methicillin-resistant *Staphylococcus aureus* by use of weekly surveillance cultures. *J. Clin. Microbiol.,* **47**, 1229–30.

Jones, M., Nielson, C., Gupta, K., Khader, K. & Evans, M. (2015) Collateral benefit of screening patients for methicillin-resistant *Staphylococcus aureus* at hospital admission: isolation of patients with multidrug-resistant gram-negative bacteria. *Am. J. Infect. Control.,* **43**, 31–34.

Khader, K., Thomas, A., Greene, T., Redd, A., Leecaster, M., Zhang, Y., Huskins, W. C. & Samore, M. H. (2017) A dynamic transmission model to evaluate the effectiveness of infection control strategies. *Open Forum Infect. Dis.*, **4**, https://doi.org/10.1093/ofid/ofw247.

Kypraios, T., O'Neill, P. D., Huang, S. S., Rifas-Shiman, S. L. & 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 Infect. Dis.,* **10**, 29.

McBryde, E. S., Pettitt, A. N., Cooper, B. S. & McElwain, D. L. S. (2007) Characterizing an outbreak of vancomycin-resistent enterococci using hidden Markov models. *J. R. Soc. Interface* **4**, 745–754.

McCallum, H., Barlow, N. & Hone, J. (2001) How should pathogen transmission be modelled? *Trends Ecol. Evol.,* **16**, 295–300.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. & Teller, A. H. (1953) Equations of state calculations by fast computing machines. *J. Phys. Chem.,* **21**, 1087–1091.

R Core Team (2015) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Shenoy, E. S., Paras, M. L., Noubary, F., Walensky, R. P. & Hooper, D. C. (2014) Natural history of colonization with methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant enterococcus (VRE): a systematic review. *BMC Infect. Dis.,* **14**, 177.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Series B,* **64**, 483–639.

Stephens, M. (2000) Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Stat.,* **28**, 40–74.

Thomas, A., Redd, A., Khader, K., Leecaster, M., Greene, T. & Samore, M. (2015) Efficient parameter estimation for models of healthcare-associated pathogen transmission in discrete and continuous time. *Math. Med. Biol.,* **32**, 79–98.

Watanabe, S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criteria in singular learning theory. *J. Mach. Learn. Res.,* **11**, 3571–3594.

Worby, C. J., Jeyaratnam, D., Robotham, J. V., Kypraios, T., O'Neill, P. D., Angelis, D. D., French, G. & Cooper, B. S. (2013) Estimating the effectiveness of isolation and decolonization measures in reducing transmission of methicillin-resistant *Staphylococcus aureus* in hospital general wards. *Am. J. Epidemiol.,* **177**, 1306–1313.

Appendix A

The Caley–Hamilton theorem states that if Q is an $n \times n$ matrix and I_n is the $n \times n$ identity matrix, then the characteristic polynomial of Q which is defined as

$$p(\lambda) = \det(\lambda I_n - Q), \tag{A.1}$$

where det denotes the determinant and λ is a scalar, solves the equation

$$p(Q) = 0. \tag{A.2}$$

It is a consequence of the Caley–Hamilton theorem that given an $n \times n$ matrix Q, its characteristic polynomial p, and an analytic function, f, we can write

$$f(x) = q(x)p(x) + r(x), \tag{A.3}$$

where q is the quotient obtained by dividing f by p, and r is the remainder polynomial which is known to have degree $\leq n$. Thus,

$$f(Q) = r(Q) = a_0 I_n + a_1 Q + \cdots + a_{n-1} Q^{n-1}. \tag{A.4}$$

Hence, provided we can evaluate the polynomial r, we will have an explicit formula for $f(Q)$. But because each eigenvalue λ of Q satisfies $p(\lambda) = 0$, then we also have $p(\lambda) = r(\lambda)$. Consequently, for each eigenvalue $\lambda_i, i = 1, 2, \ldots, n$, we have

$$f(\lambda_i) = r(\lambda_i) = a_0 + a_1 \lambda_i + \cdots + a_{n-1} \lambda_i^{n-1}. \tag{A.5}$$

So in order to determine r, we simply need to know the coefficients $a_0, \ldots, a_{n-1}$, which can be done by solving a system of linear equations.

To compute the transition probabilities for the 3-state outpatient Markov process, we need to evaluate $f(Q) = e^{tQ}$, where Q is the transition rate matrix of the process, and is given by (11). From (A.4), we have

$$e^{tQ} = a_0 I_3 + a_1 Q + a_2 Q^2. \tag{A.6}$$

Q has three eigenvalues, $\lambda_1 = 0$, and the other two which are complex conjugate are given by

$$\lambda_{2,3} = \frac{-(\kappa + \mu + \nu) \pm \sqrt{\kappa^2 + \mu^2 + \nu^2 - 2\kappa\mu - 2\kappa\nu - 2\mu\nu}}{2}. \tag{A.7}$$

The coefficients in (6) solve the system of equations given in (5). In particular, the system of equations necessary to obtain the coefficients to r is given by

$$\begin{cases} 1 = a_0 \\ e^{\lambda_2 t} = a_0 + a_1 \lambda_2 + a_2 \lambda_2^2 \\ e^{\lambda_3 t} = a_0 + a_1 \lambda_3 + a_2 \lambda_3^2 \end{cases} \tag{A.8}$$

from which we can readily solve for a_0, a_1 and a_2 to get $a_0 = 1$,

$$a_1 = \frac{\lambda_2^2(e^{\lambda_3 t} - 1) - \lambda_3^2(e^{\lambda_2 t} - 1)}{(\lambda_2 - \lambda_3)\lambda_2\lambda_3} \tag{A.9}$$

and

$$a_2 = \frac{\lambda_3(e^{\lambda_2 t} - 1) - \lambda_2(e^{\lambda_3 t} - 1)}{(\lambda_2 - \lambda_3)\lambda_2\lambda_3} \tag{A.10}$$

Note that, although a_1 and a_2 are necessarily real, λ_2 and λ_3 are in general complex numbers and our implementation of these computations use the standard complex variable type and functions of C++.