JOURNAL OF THE ROYAL SOCIETY Interface

Characterizing an outbreak of vancomycin-resistant enterococci using hidden Markov models

E.S McBryde, A.N Pettitt, B.S Cooper and D.L.S McElwain

J. R. Soc. Interface 2007 **4**, 745-754
doi: 10.1098/rsif.2007.0224

JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2007) **4**, 745–754
doi:10.1098/rsif.2007.0224
Published online 13 March 2007

Characterizing an outbreak of vancomycin-resistant enterococci using hidden Markov models

E. S. McBryde[1,*], A. N. Pettitt[1], B. S. Cooper[2] and D. L. S. McElwain[1]

[1]*School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001, Australia*
[2]*Modelling and Bioinformatics Department, Centre for Infections Health Protection Agency, 61 Colindale Avenue, London NW9 5EQ, UK*

Background. Antibiotic-resistant nosocomial pathogens can arise in epidemic clusters or sporadically. Genotyping is commonly used to distinguish epidemic from sporadic vancomycin-resistant enterococci (VRE). We compare this to a statistical method to determine the transmission characteristics of VRE.

Methods and findings. A structured continuous-time hidden Markov model (HMM) was developed. The hidden states were the number of VRE-colonized patients (both detected and undetected). The input for this study was weekly point-prevalence data; 157 weeks of VRE prevalence. We estimated two parameters: one to quantify the cross-transmission of VRE and the other to quantify the level of VRE colonization from sporadic sources. We compared the results to those obtained by concomitant genotyping and phenotyping.

We estimated that 89% of transmissions were due to ward cross-transmission while 11% were sporadic. Genotyping found that 90% had identical glycopeptide resistance genes and 84% were identical or nearly identical on pulsed-field gel electrophoresis (PFGE).

There was some evidence, based on model selection criteria, that the cross-transmission parameter changed throughout the study period. The model that allowed for a change in transmission just prior to the outbreak and again at the peak of the outbreak was superior to other models. This model estimated that cross-transmission increased at week 120 and declined after week 135, coinciding with environmental decontamination.

Significance. We found that HMMs can be applied to serial prevalence data to estimate the characteristics of acquisition of nosocomial pathogens and distinguish between epidemic and sporadic acquisition. This model was able to estimate transmission parameters despite imperfect detection of the organism. The results of this model were validated against PFGE and glycopeptide resistance genotype data and produced very similar results. Additionally, HMMs can provide information about unobserved events such as undetected colonization.

Keywords: HMM; nosocomial pathogens; genotyping; statistical modelling; VRE

1. INTRODUCTION

There has been an alarming worldwide increase in the rate of infection from vancomycin-resistant enterococci (VRE) in the last 15 years (Murray 2006). Enterococci are part of the normal gastrointestinal flora and VRE colonization often is asymptomatic and undetected. However, in patients with compromised immune systems and breached integument, enterococci can become pathogenic, causing, for example, urinary tract infection, bacteraemia and endocarditis. Large teaching hospitals and intensive care units (ICUs) have the highest rate of infection with VRE (Weinstein 2005). Infection with enterococci harbouring a vancomycin resistance gene is associated with higher mortality (Lodise *et al.* 2002) and many strains of VRE are resistant to all known antibiotics.

Acquisitions of VRE colonization can be grouped broadly into those that come from cross-transmission within the ward which we call *transmitted*, and VRE that comes from other sources which we call *sporadic*. Ward transmission of multi-resistant organisms is believed to be predominantly from patient to patient via the transiently contaminated hands of health care workers (Boyce 2001). The sources of sporadic VRE include patients' gastrointestinal tract, prior colonization with VRE and transmission from outside the ward. The presence of VRE on admission is often initially not detected owing to infrequent swabbing, poor sensitivity of swabs or undetectable quantities of organism. VRE may exist in subdetectable numbers in human gut so that exposure of patients to antibiotics which facilitate VRE growth (Donskey *et al.* 2002) may lead to an

*Author for correspondence (e.mcbryde@qut.edu.au).

Received 17 January 2007
Accepted 9 February 2007

apparently new case of VRE. VRE is also known to spread from other hospital wards via patient and staff movements (Trick *et al.* 1999).

To select the most appropriate infection control interventions, one needs to be able to estimate how much of the new acquisition is transmitted and how much is sporadic. Restricting antibiotic exposure is thought to control sporadic VRE, by reducing selection pressure in patients' endogenous flora, while hand hygiene, cohorting, patient isolation and limiting admission of colonized patients are thought to impact on transmitted VRE.

Outbreak investigation often involves time intensive methods to characterize the mode of VRE acquisition. Genotyping techniques such as pulsed-field gel electrophoresis (PFGE), distinguish clonal outbreaks, which are presumed to be due to transmitted VRE, from multiple new strain introductions, which are presumed to be due to sporadic VRE. There are occasions when this technique breaks down, when horizontal transfer of the resistance gene, vanA or vanB, can lead to several different genotypes being detected when in fact a single transposon is being transmitted (Suppola *et al.* 1999; Bradley *et al.* 2002; Weinstein 2005).

Attempts have been made to distinguish between these two processes of colonization based on statistical analysis of surveillance data. Pelupessy *et al.* (2002) used a Markov model, without hidden states, to estimate transmission parameters; finding estimates were similar to those using full event data and genotyping (PFGE). Cooper & Lipsitch (2004) used structured and unstructured hidden Markov models (HMMs) to describe infection incidence time-series data and to estimate transmission parameters. Collinearity between parameter estimates, failure of convergence and computational difficulties were identified as potential problems using HMMs for sparse data such as is typically found in time-series infection control data. Forrester & Pettitt (2005) compared background rates with cross-transmission rates of methicillin-resistance *Staphylococcus aureus*, finding background rates were larger than cross-transmission rates.

Estimating transmission coefficients using hospital infection control data has a number of challenges. There are unobserved processes occurring; the time of new acquisition of colonization is not observed. Additionally, when relying on routine swabs to determine the number of colonized patients, the sensitivity of swabs is less than 100%.

This study uses an epidemic model structure to characterize transmission of VRE during an outbreak at an 800 bed Australian teaching hospital. The current paper extends the work by Pelupessy *et al.* (2002) by estimating epidemiological parameters in the presence of suboptimal swab sensitivities. It also allows for the fact that new colonizations are not immediately detectable. We use an HMM structure to estimate transmission in the face of incomplete datasets and unobserved events. This framework distinguishes between rates of transmitted and sporadic VRE acquisitions. This study also considers that the transmission rates may change over time. Section 2.1 describes the data used to estimate VRE epidemic determinants. Section 2.4 describes the model of VRE transmission, while §2.5 describes the HMM and the methodology behind it. Section 3 gives the results of the parameter estimates, comparison of model estimates and genotyping data and model selection.

2. METHODS

2.1. Description of outbreak and infection control interventions

VRE was first isolated at the Princess Alexandra Hospital, Brisbane, Australia in October 1996 and a VRE screening programme commenced in January 1997, the beginning of the data collection period for this study. Data used in this study are VRE colonization data from the ICU, renal and infectious diseases units. VRE colonized patients and were identified by clinical isolates, weekly routine screening and contact tracing swabs. Infection control interventions introduced from the start of the study period were restriction of vancomycin and third-generation cephalosporin use and isolation of colonized patients. From week 125 of this study, infection control teams were aware of an increased prevalence of VRE and further measures were taken. Dedicated equipment was used in patient rooms and patients were cohorted. VRE patients requiring haemodialysis used a dialysis facility within the infection control unit. Medical and nursing staff wore disposable aprons and latex gloves for patient contacts. An environmental audit was performed in August 1999, approximately week 135 of the study period and an aggressive cleaning programme was instituted (Bartley *et al.* 2001).

2.2. Serial surveillance data used for statistical analysis

Input data for the statistical model in this study were:

— Weekly prevalence data for VRE colonization.
— Mean length of stay of colonized patients: 15 days. This was calculated as the time from first identification of colonization to discharge.
— The total number of beds in the wards, $N=68$.

The data were collected from 1 January 1997 to 31 December 1999. The weekly prevalence data are shown in figure 1.

2.3. Data used for cluster analysis

Microbiological and clinical data were collected, including admission dates and discharge dates of VRE colonized patients, as well as the date of first positive isolate. Additionally, we had information on the colonization status on admission of three of the patients transferred from other hospitals. Genotype data, both PFGE and glycopeptide resistance genotyping, were compared with the results of the statistical analysis as part of the study validation. Presumptive VRE colonies were identified using standard techniques. Speciation (distinguishing *Enterococcus faecium* and *Enterococcus faecalis*) was initially achieved by carbohydrate fermentation reactions of arabinose, mannose and raffinose then confirmed by a multiplex PCR assay based on

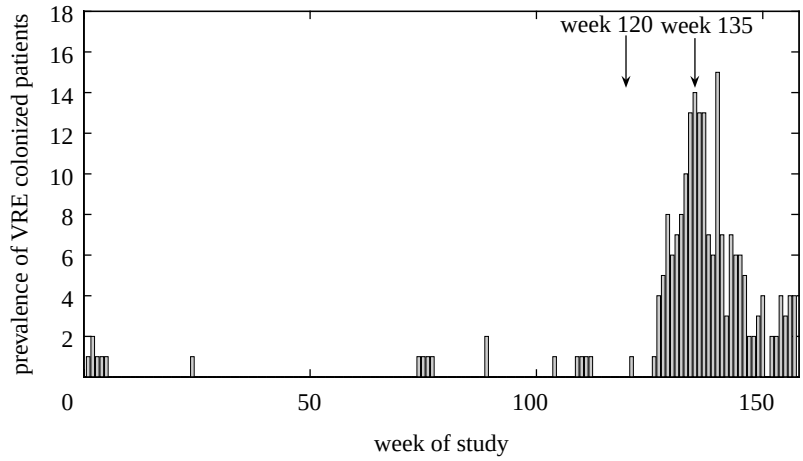

Figure 1. Prevalence data for VRE over 157 weeks. Arrows show times in which changes in transmission rates may have taken place.

specific detection of genes encoding D-alanine: D-alanine ligases (Bartley *et al.* 2001). VRE phenotype was identified based on vancomycin and teichoplanin mean inhibitory concentrations using the *E*-test method. This presumptively distinguishes vanA VRE, resistant to both vancomycin and teichoplanin, from vanB VRE, resistant to vancomycin but sensitive to teichoplanin. This phenotype result was confirmed by glycopeptide resistance genotyping, achieved through a modified multiplex PCR assay, described in detail by Bartley *et al.* (2001).

In the study on this outbreak by Bartley *et al.* (2001), isolates were also characterized using PFGE. Electrophoretic band patterns were analysed according to the criteria established by Tenover *et al.* (1995). Computer comparison using GEL COMPAR v. 4.1 (Applied Maths Kortrijk, Belgium) was based on the algorithm of the unweighted pair group method for arithmetic averages and using the Dice coefficient with 1.5% band tolerance (Bartley *et al.* 2001). This information was used to estimate the proportion of isolates that were from the same strain.

2.4. Model of transmission

We based our ward transmission model on the Susceptible-infected model with migration, described by Bailey (1975). Modified versions of this model have been used previously to analyse nosocomial transmission data (Pelupessy *et al.* 2002; Cooper & Lipsitch 2004; Forrester & Pettitt 2005).

A schematic of the model is shown in figure 2. The rate of cross-transmission of VRE colonization (per colonized per susceptible patient per day) is denoted by β. It is assumed that the ward is of fixed size, N, hence the number of uncolonized patients is $N-C$. Colonized patients are assumed to remain colonized for their entire hospital stay, therefore, transition from colonized to uncolonized occurs via discharge of a colonized patient and replacement with an uncolonized patient, which occurs at a rate μ. Duration of stay of colonized patients was available from the dataset. Acquisition of VRE, that is transmitted, is described by the mass-action term, $\beta C(N-C)$. VRE acquisition, that is sporadic, can arise

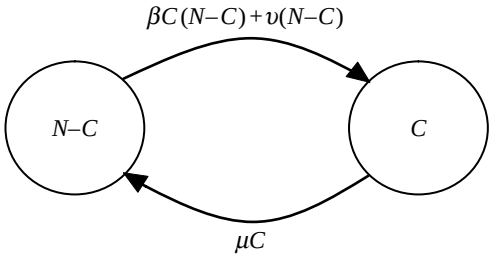

Figure 2. The transmission of bacterial pathogens in the hospital ward.

through ward admission of a colonized patient or any other process that is not related to the number of colonized patients, and occurs at a rate, $\nu(N-C)$. Each of the processes that lead to sporadic acquisition (for example, prior colonization or colonization from out-of-ward sources, endogenous gastrointestinal colonization) can reasonably be assumed to be independent of the number of colonized patients in the ward.

The probability of a change in the number of colonized patients, C, in a short time period, h, is given by

$$\begin{aligned}
&\Pr[C(t+h) = i+1 | C(t) = i] \\
&\quad = \beta i(N-i)h + \nu(N-i)h + o(h), \\
&\Pr[C(t+h) = i-1 | C(t) = i] = \mu i h + o(h), \\
&\Pr[C(t+h) = i | C(t) = i] \\
&\quad = 1 - \beta i(N-i)h - \nu(N-i)h - \mu i h + o(h), \\
&\Pr[C(t+h) = j (j \neq i-1, i, i+1) | C(t) = i] = o(h).
\end{aligned} \tag{2.1}$$

The number of colonized patients in the ward, $C(t)$, forms a Markov process on state space 0, ..., N, where N is the number of patients on the ward. Reflecting boundaries occur at states $i=0$ and $i=N$, provided $\nu>0$, otherwise 0 is an absorbing state, and provided $\mu>0$, otherwise N is an absorbing state.

2.5. Hidden Markov model

We aim to estimate parameters associated with sporadic colonization, ν, and the colonization caused by ward *transmission*, β, using the structured HMM illustrated in figure 3.

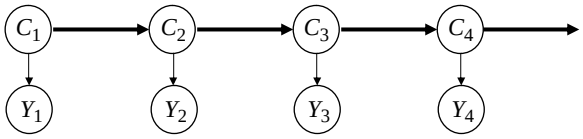

Figure 3. Hidden Markov model. Here, C represents the number of colonized patients in the ward (detected or undetected), Y represents the number of patients detected at each time point. The horizontal arrows represent the transition from one state to the next, and the vertical arrows represent the relationship between the hidden state and the corresponding observation.

Our HMM consists of: observations, $\boldsymbol{Y}$, the number of patients detected at each time point; underlying hidden states, $\boldsymbol{C}$, the number of colonized patients in the ward; a transition model linking each hidden state with its adjacent states, represented by horizontal lines in figure 3; an observation model linking the data with the hidden state, represented by the vertical lines in figure 3. There is one hidden state for each observation, denoted $C_1, C_2, \ldots, C_n$.

The full conditional probability of any node depends only on neighbouring nodes to which it is connected directly. The observation component of the HMM, denoted by $\boldsymbol{Y}$, consists of 157 data inputs of weekly VRE prevalence taken over 3 years and the vector of time points, $\boldsymbol{t}=t_1, \ldots, t_n$, corresponding to each observation time. The vector $\boldsymbol{C}$ consists of the $n=157$ hidden states. The transition probability matrix, giving the relationship between the hidden states, is described in appendix A. The observation model, giving the relationship between the observed and hidden states, is described in §2.6.

The parameters used in the model are given in table 1.

2.5.1. Model assumptions. The model makes the following assumptions

(i) The ward is of fixed size, N.
(ii) The model parameters are time invariant (this assumption is relaxed later in the study).
(iii) Each colonized patient remains colonized for the remainder of their stay.
(iv) Each observation of patients not known to be colonized is conditionally independent given the corresponding hidden state.
(v) The hidden states follow a first order time homogenous Markov process, that is

$$\begin{aligned}&\Pr(C(t_k)|C(t_1), \ldots, C(t_{k-1}))\\ &\quad = \Pr(C(t_k)|C(t_{k-1}))\\ &\quad = \Pr(C(t_k - t_{k-1})|C(0)).\end{aligned}$$

(vi) Homogenous mixing of patients takes place.
(vii) Uncolonized patients are identical with respect to susceptibility to colonization.
(viii) Colonized patients are identical with respect to transmission of VRE.
(ix) Time from colonization to discharge is exponentially distributed. Review of patient histories confirms that this is approximately the case.

These assumptions are discussed in §4.

Table 1. Parameters used in the model. Fitted values are discussed in §3.

parameter	symbol	value	source
number of patients	N	68	directly from dataset
removal rate of colonized patient	μ	$1/15\ \text{day}^{-1}$	directly from dataset
transmission rate	β	1.0×10^{-3}	fitted using HMM
sporadic acquisition rate	ν	2.0×10^{-4}	fitted using HMM
detection probability	D	0.58–0.97	literature review

2.6. Observation model

The probability of being known to be colonized (and therefore being included in the prevalence data) given that a patient is colonized, d, was unknown. Literature sources regarding the sensitivity of rectal swabs in detecting VRE were used to develop an expression for the uncertainty in this parameter. Estimates of the sensitivity of a rectal swab for VRE range from 0.58 (D'Agata *et al.* 2002) to 0.97 (Reisner *et al.* 2000) with values in between (Lemmen *et al.* 2001; Trick *et al.* 2004). We allowed for the uncertainty regarding the detection by assigning a uniform $[0.58, 0.97]$ prior distribution to d. The probability of detection at a given prevalence check, d, used in this study was patient related rather than simply swab related. If a patient was known from previous swabs to be colonized, the patient was automatically detected, thus d would be expected to be at least as high as the sensitivity of a single swab. The observation model assumed that each week's observed prevalence is independent of the previous week's observed prevalence, given the underlying true prevalence. This is an approximation as the true detection is the known colonized patients from the previous week and the new colonizations from the current week.

The probability relationship between the states and the data is described by the binomial distribution $Y_k \sim \text{Bin}(C_k, d)$, where Y_k is the kth observed colonization prevalence and C_k is the actual number of colonized patients, the hidden state, at time t_k. This assumes that the probability, d, remains constant over the study period (for each iteration) and the probability of detection of each colonized patient is independent of the number of other colonized patients.

Alternative observation models with greater dispersion could have been used. For example, the Poisson or negative binomial distribution could have been chosen, had we been dealing with incidence rather than prevalence data. We chose the binomial distribution because it has a sound probabilistic basis (assuming fixed detection) and, unlike the Poisson, ensures that the hidden state (number colonized) is always larger than the observation (number detected), a necessary result when using prevalence data.

2.7. Bayesian framework

The parameters for transmitted VRE, β and sporadic VRE, ν were estimated using a Bayesian framework. Let $\boldsymbol{\theta}_{\mathrm{p}}=\{\beta, \nu, d\}$ be the vector of model parameters. Baum *et al.*'s (1970) recursion formula, summarized in appendix B, was used to determine the likelihood of the data, $l(\boldsymbol{Y}|\boldsymbol{\theta}_{\mathrm{p}})$. Uniform $U[0, 0.1]$ priors were assigned to β and ν, because little was known about these parameters other than that negative values or values higher than 0.1 were completely implausible. The posterior probability distributions

$$\Pr(\boldsymbol{\theta}_{\mathrm{p}}|\boldsymbol{Y}) \propto \pi(\boldsymbol{\theta}) l(\boldsymbol{Y}|\boldsymbol{\theta}_{\mathrm{p}}), \tag{2.2}$$

were estimated using a Monte-Carlo Markov chain algorithm, described in appendix C.

The Bayesian framework can provide estimates (and full posterior probability density) of any function of model parameters including functions which depend upon knowledge of hidden states. Let $\boldsymbol{\theta}_{\mathrm{h}}$ be the vector of n inferred hidden states $C_1, \ldots, C_n$ and let $\boldsymbol{\theta}=\{\boldsymbol{\theta}_{\mathrm{p}}, \boldsymbol{\theta}_{\mathrm{h}}\}$. The expected number of within-ward transmissions for the week, following week k is $\beta C_k(N-C_k)$, while the total number of transmissions is $\beta C_k(N-C_k) + \nu(N-C_k)$. The expected proportion of VRE acquisitions due to ward transmission over the time of the study, $f(\boldsymbol{\theta})$, is approximated by

$$f(\boldsymbol{\theta}) = \frac{\sum_{k=1}^{n} \beta C_k(N-C_k)}{\sum_{k=1}^{N} \beta C_k(N-C_k) + \nu(N-C_k)}. \tag{2.3}$$

We evaluate the expectation, $E[f(\boldsymbol{\theta})|\boldsymbol{Y}]$, by drawing samples $\boldsymbol{\theta}_k$, $k=1, \ldots, m$ from $p(\boldsymbol{\theta}|\boldsymbol{Y})$ and using the approximation of Gilks *et al.* (1996, ch. 1)

$$E[f(\boldsymbol{\theta})|Y] \approx \frac{1}{m}\sum_{k=1}^{m} f(\boldsymbol{\theta}_k). \tag{2.4}$$

The algorithm for this Monte-Carlo integration is given in appendix C.

2.8. Comparison of cluster analysis results using genotyping with statistical analysis

A genotyping study was performed on the VRE isolates by Bartley *et al.* (2001). Of the 49 isolates available for analysis, 44 were found to be *E. faecium* vanA using glycopeptide resistance genotyping. The estimated number of isolates having identical or closely related patterns on PFGE using the criteria of Tenover *et al.* (1995) was 41 of 49.

2.8.1. Cluster analysis based on genotypic relatedness. We compared the proportion of 'identical isolates' (presumed to be part of a cluster) with the estimated proportion of transmitted VRE derived from the HMM and prevalence data. The posterior probability distribution of the proportion of VRE cases that are identical by genotype can readily be derived using a Bayesian framework and conjugate prior distribution (Gelman *et al.* 2004). Denote the parameter of interest, the proportion of VRE acquisitions that are identical, by p. Assume the form Beta(1, 1) for the prior distribution for the proportion; this is the same as the uniform[0, 1] prior. The probability of the data is given by the binomial Bin(a; $(a+b)$, p), where a is the number of identical isolates and b is the number of non-identical isolates, as detected by the laboratory methods. The posterior probability density of p is Beta($1+a$, $1+b$).

3. RESULTS

3.1. Transmission parameter estimation

The estimated value for the transmission coefficient, β was 10×10^{-4} (CI_{95} 7.9×10^{-4}, 13×10^{-4}) and the sporadic acquisition rate ν was 2.0×10^{-4} (CI_{95} 0.85×10^{-4}, 3.8×10^{-4}). The coefficient of correlation between β and ν was estimated to be -0.24. These results were obtained using a Markov chain Monte-Carlo algorithm with a burn-in period of 50 000 as described in appendix C.

The basic reproduction ratio, R_0, is 'the average number of persons directly infected by an infectious case during its entire infectious period, after entering a totally susceptible population' (Giesecke 1994). In this model it can be shown to be $R_0=\beta N/\mu$. This formula for R_0 is an approximation as there is a finite population in this setting. The basic reproduction ratio is estimated to be 1.07 (CI_{95} 0.78–1.34).

The mean value for the estimated detection rate, d, was 0.75 with a 95% credible interval of 0.59–0.93.

3.2. Comparison of statistical model and genotyping data

The proportion of VRE acquisitions due to transmission, was estimated to be 89% (CI_{95}=78–95%), using Bayesian inference applied to the HMM structure. This compares with 84% (41/49) of isolates observed to be identical or nearly identical using PFGE genotyping and 90% (44/49) using glycopeptide resistance genotyping. The posterior distribution of the estimated proportion of colonizations due to ward transmission compared with those found to be identical by glycopeptide resistance genotype and PFGE methods are displayed in figure 4.

3.3. Sensitivity analysis

The length of stay following colonization could be greater than the estimated 15 days because acquisition could have preceded detection. Conversely, the length of stay could have been less than 15 days because undetected colonized patients are likely to have shorter stays than detected colonized patients. We therefore performed a sensitivity analysis on the discharge rate parameter, μ. We took upper and lower values for μ which we believed at the extreme ends of plausibility. We then repeated the estimation of the proportion of VRE acquisitions due to within-ward transmission. Results are given in table 2.

Table 2 shows there is a small change in the estimate for large changes in the discharge parameter, μ.

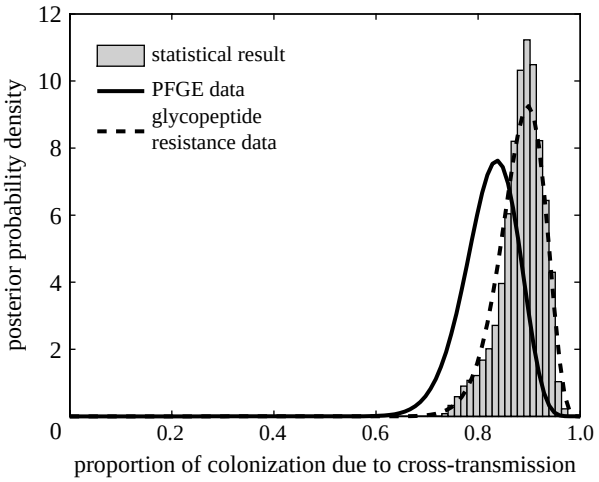

Figure 4. Posterior distribution of proportion of VRE acquisitions that are due to ward transmission. The histogram gives the posterior distribution from the Bayesian analysis of the HMM, the solid curve gives the posterior distribution based on the observed proportion of identical strains using PFGE genotype data and the broken line gives the posterior distribution based on observed proportion of identical strains using glycopeptide resistance phenotype and genotype data (Bartley *et al.* 2001).

Table 2. Analysis of sensitivity of the model outcome to changes in the discharge rate, μ.

μ	estimated proportion (%)
1/10	91.4
1/15	89
1/20	86.5

3.4. Model selection

The values of the deviance information criterion (DIC) were used to assess the optimum model to fit the data (Gelman *et al.* 2004). Results are given in table 3.

Several models were explored. Setting either β or ν to zero led to much higher values for the DIC, giving substantial statistical support to a mixed model, in which VRE colonization arose both from cross-transmission in the ward and sporadically. The model in which β changed after week 120 was a superior fit to the model with time-invariant parameters. Allowing for a further change in β after week 135 provided the best fit of those models investigated. The effective number of parameters in a latent variable model depends on the collinearity of the parameters and the influence of the latent variables.

4. DISCUSSION

The aim of this study was to characterize transmission of VRE using statistical methods and simple serial surveillance data. We included a term for sporadic colonization because we believe that new acquisitions of VRE could occur through means other than within-ward patient to patient cross-transmission. Sources of sporadic colonization have been labelled in the past as endogenous, spontaneous (Pelupessy *et al.* 2002) or background (Forrester & Pettitt 2005). Our statistical methods were designed to quantify the rates of sporadic and cross-transmitted VRE. Previous attempts have encountered difficulties especially with identifiability of variables (Cooper & Lipsitch 2004).

Full patient histories, PFGE and glycopeptide resistance genotype data were used for validation but were not included in the statistical analysis in this study. Estimates of the proportion of VRE resulting from cross-transmission based on statistical methods (HMMs) in this study were very similar to those based on vancomycin resistance genotype data.

The proportion of clustered isolates based on PFGE analysis was lower than both the vancomycin resistance genotype data and the statistical analysis. This could be due to horizontal transfer of resistance gene to new strains of enterococci, which has been reported previously (Suppola *et al.* 1999; Weinstein 2005). If horizontal transfer of resistance genes occurs during an outbreak, cross-transmitted strains have identical glycopeptide resistance genotypes but different PFGE patterns, hence PFGE underestimates clustering.

Using a structured HMM, one can estimate the hidden states behind the data, the number of patients colonized on the ward (both detected and undetected). We estimated the basic reproduction ratio to be close to unity, the threshold value that could lead to endemic VRE. We were able to make estimates of transmission in the face of imperfect datasets in which transmission times and patient histories were unknown and swab sensitivity was considerably less than 100%. This approach is similar to that of Cooper & Lipsitch (2004), who observed monthly infection *incidence* and assumed a Poisson relationship with the number colonized, the hidden state. The current study avoids the ambiguity of the relationship between the observations and the hidden state using *prevalence* (observed number detected at time points) which relates directly to the hidden state, the number colonized, through a binomial relationship.

For simplicity, this study assumed homogenous mixing of staff and patients. Future studies could extend this model to include ward coupling, however, dividing the data to incorporate ward structure would lead to reduced precision in parameter estimates and increased model complexity. We incorporated this uncertainty into our parameter estimates and model conclusions were robust to changes in its value.

The time to discharge was estimated by taking the mean time from first identification of colonized patients to discharge (15 days). The discharge rate was taken as the reciprocal of the mean time to colonization. This assumes that the time to discharge was exponentially distributed which was indeed approximately the case for those known to be colonized. The true time to discharge of colonized patients could have been longer than estimated in this study if patients were colonized for significant time-intervals prior to detection or they could have been shorter if a substantial number of the undetected colonized patients had shorter durations of stay. We performed a sensitivity analysis on the discharge rate parameter, μ, and found large changes in μ ($\pm 33\%$) resulted in small changes in the estimates

Table 3. Comparison of different models using the deviance information criterion. P_d: effective number of parameters.

model	estimate of β (95% CI)$\times 10^{-4}$	estimate of ν (95% CI)$\times 10^{-4}$	DIC	P_d
one value for ν and three values for β with change points at the end of week 120 and 135	$\beta_1$3.4(0.28–8.8) $\beta_2$15.3(13.5–17.1) $\beta_3$10.9(7.1–13.0)	2.2(0.96–4.0)	251	4.0
one value for ν and two values for β with change point at the end of week 120	$\beta_1$3.4(0.28–8.7) $\beta_2$11.9(10.2–13.5)	2.2(0.96–4.0)	253	2.3
one value for ν and one value for β	10(7.9–13)	2.0(0.85–3.8)	261	2.6
one value for ν and two values for β with change point at the end of week 135	$\beta_1$11(7.6–14.6) $\beta_2$9.6(7.9–11.4)	2.0(0.88–3.7)	261	2.6
$\beta=0$ and one value for ν	0	9.7(7.7–11.7)	393	1.2
$\nu=0$ and one value for β	8.7(6.9–10.1)	0	531	1.5

of proportion of patients colonized within the ward and is therefore unlikely to have influenced the conclusions of this study.

The model presented in this study postulated that VRE acquisition arose from both cross-transmission and sporadic sources. Model comparison techniques found this model to be a far superior fit to the data compared with models which relied on either cross-transmission or sporadic sources of VRE acquisition alone, strongly supporting that both modes of acquisition were taking place.

We investigated changes in transmission over time using a structured epidemic model. Model comparison showed that there was evidence supporting the conclusion that there was an increase in cross-transmission just prior to the outbreak. There was limited evidence that the cross-transmission rate reduced after the epidemic peak at week 135, coinciding with the environmental cleaning intervention. Future studies using larger surveillance datasets could extend the methodology presented to consider more models in which parameters are time-dependent. One approach to this would be to use the reversible jump Monte-Carlo Markov chain method (Green 1995) or the birth–death Markov process model (Stephens 2000).

Inaccuracies in PFGE cluster analysis can arise from the horizontal transfer of resistance genes. Glycopeptide resistance genotype analyses are not subject to inaccuracies due to gene transfer but cannot distinguish different strains that might all be of the same resistance genotype. Statistical methods are not subject to these problems and have the additional advantage that they are not resource intensive. It is interesting to speculate whether they also have the potential to be used in real time, within a control-chart outbreak alert system.

The model presented here can be used to model the transmission of other bacterial pathogens in small scale settings of healthcare institutions, such as methicillin-resistant *Staphylococcus aureus*, extended spectrum beta-lactamase producing and other multi-resistant Gram-negative pathogens.

This work was partially supported by a grant under the Australian Research Council Linkage Scheme (LP0347112) and NHMRC scholarship number 290541. The authors would like to thank Dr Mike Whitby for providing data and Dr Paul Bartley for helpful comments. The authors would like to thank the anonymous reviewers for their constructive comments.

APPENDIX A. CONSTRUCTING A TRANSITION PROBABILITY MATRIX

Following the theory of Cox & Miller (1965), we developed a transition probability matrix, $\boldsymbol{\Gamma}_{(t_k-t_{k-1})}$. The ijth element of $\boldsymbol{\Gamma}_{(t_k-t_{k-1})}$ gives the probability of having j colonized patients on the ward at time t_k, given that there were i colonized patients on the ward at time t_{k-1}.

To construct the transition probability matrix for an arbitrary time-interval, first we developed a discrete time transition probability matrix, $\boldsymbol{A}$, for a small time-interval, h. Let $\boldsymbol{A}$ be the matrix in which the ijth element is given by $\Pr(C(t+h)=j|C(t)=i)$. A is given using the system of equation (2.1). Here, i and j are the number of patients colonized in the ward and can take on values 0, …, N.

Let $\boldsymbol{p}(t)$ be the $(N+1)$ vector of probabilities of the number colonized at time t. The generator matrix, $\boldsymbol{G}$ is a square, $(N+1)\times(N+1)$, matrix that has the property that

$$\frac{\mathrm{d}\boldsymbol{p}(t)}{\mathrm{d}t} = \boldsymbol{G}\boldsymbol{p}(t). \tag{A 1}$$

The ijth element of the generator matrix, $\boldsymbol{G}$, is the instantaneous rate of change of probability of being in state j, given a beginning in state i. Then $\boldsymbol{G}$ is given by

$$\boldsymbol{G} = \lim_{h\to 0}\frac{1}{h}(\boldsymbol{A}-\boldsymbol{I}), \tag{A 2}$$

where $\boldsymbol{I}$ is the identity matrix.

Following from expression (A 2), we have

$$\boldsymbol{p}(t_{k+1}) = \boldsymbol{p}(t_k)\mathrm{e}^{(t_{k+1}-t_k)\boldsymbol{G}}, \tag{A 3}$$

in general. Specifically, after a time-interval t_k–t_{k-1}, the probability of being in state j having begun in state i is the ijth element of the transition probability matrix, given by

$$\boldsymbol{\Gamma}_{(t_k-t_{k-1})_{ij}} = \Pr(C_k=j|C_{k-1}=i) = (\mathrm{e}^{(t_k-t_{k-1})\boldsymbol{G}})_{ij}. \tag{A 4}$$

Cox & Miller (1965, ch. 4.5) and MacDonald & Zucchini (1997) give an expanded explanation. The matrix exponential $\mathrm{e}^{(t_k-t_{k-1})\boldsymbol{G}}$ was calculated using the MATLAB 'expm' function.

APPENDIX B. LIKELIHOOD COMPUTATION

The probability of the full dataset and a particular sequence of hidden states, C_1, C_2, ..., C_n is given by

$$\Pr(Y_1, \ldots, Y_n, C_1, \ldots, C_n|\beta, \nu) = \Pr(C_1)\Pr(Y_1|C_1)\prod_{k=2}^{n} \boldsymbol{\Gamma}_{C_{k-1}C_k}\Pr(Y_k|C_k), \quad \text{(B 1)}$$

with $\boldsymbol{\Gamma}_{C_{k-1}C_k}$ as defined in appendix A.

The likelihood calculation of this single permutation of hidden states requires $2n$ computations even after the matrix exponential has been evaluated. The full likelihood of the data over all the states is

$$\Pr(Y_1, \ldots, Y_n|\beta, \nu) = \sum_{C_1=1}^{N+1}, \ldots, \sum_{C_n=1}^{N+1} \Pr(Y_1, \ldots, Y_n, C_1, \ldots, C_n|\beta, \nu), \quad \text{(B 2)}$$

which requires $2n(N+1)^n$ computations for one likelihood evaluation (Le Strat & Carrat 1999). This intractable calculation (with $n=157$ and $N=68$) can be simplified using Baum's recursion technique (Baum *et al.* 1970) as shown below.

The forward recursion involves simplifying the likelihood computations by considering a partial observation sequence and a single state sequence. Let $\phi_k(i)$ be the probability of the partial observation sequence (Y_1, Y_2, ..., Y_k) produced by all possible state sequences that end in state i. The probability is given by

$$\phi_k(i) = L(Y_1, \ldots, Y_k, C_k = i|\nu, \beta), \quad k \leq n. \quad \text{(B 3)}$$

Let $\boldsymbol{\delta}$ be the (size $N+1$) vector of probabilities of the first state, ($\boldsymbol{\delta}_i = \Pr(C_1 = i)$). In the forward recursion method of likelihood computation, the value of $\boldsymbol{\delta}$ needs to be determined in the absence of data. The stationary distribution of the transition matrix can be used for this (MacDonald & Zucchini 1997). The probability of the first state and first observation, Y_1, is given by

$$\phi_1(i) = \boldsymbol{\delta}_i \Pr(Y_1|C_1 = i). \quad \text{(B 4)}$$

The forward recursion formula is then applied. We multiply every state probability, $\phi_{k-1}(i)$, by the transition probability $\boldsymbol{\Gamma}_{ij}$ and by the probability of the kth data point, given the hidden state j. This results in a vector of probabilities which is then summed to determine $\phi_k(j)$. Thus, the probability of subsequent states is given by

$$\phi_k(j) = \left[\sum_{i=0}^{N} \phi_{k-1}(i)\boldsymbol{\Gamma}_{ij}\right] \Pr(Y_k|C_k = j). \quad \text{(B 5)}$$

At each step in the forward recursion, the procedure can be terminated and the probability of the partial observation sequence is determined by

$$\Pr(Y_1, \ldots, Y_k|\nu, \beta) = \sum_{i=0}^{N} \phi_k(i). \quad \text{(B 6)}$$

The likelihood of the data can then be determined by

$$\Pr(Y_1, \ldots, Y_n|\beta, \nu) = \sum_{i=0}^{N} \phi_n(i). \quad \text{(B 7)}$$

See Petrushin (2000) for a detailed discussion of the forward and backward recursion formulae.

APPENDIX C. MONTE-CARLO MARKOV CHAIN ALGORITHM

The algorithm for this Monte-Carlo integration used to estimate the proportion of VRE acquisitions due to ward cross-transmission, $f(\boldsymbol{\theta})$, is given below.

The MCMC algorithm has the following steps:

(i) Assume the prior probability for β and ν, to be ($U[0, 0.1]$). These priors were used as little prior information was known except that negative values and values greater than 0.1 are completely implausible.

(ii) Initialize β, ν and d. Different initial values were chosen from ($\beta = 10^{-5}$ to $\beta = 10^{-2}$, from $\nu = 10^{-5}$ to $\nu = 10^{-2}$) and from $d = 0.58$ to $d = 0.97$.

(iii) Assign the prior probability of the hidden states. A discrete uniform distribution on (0, ..., N) was used.

(iv) Initialize each hidden state using its corresponding observation and the (binomial) observation model $Y_k \sim \mathrm{Bin}(C_k, d)$.

(v) Determine the probability of the data and sequence of hidden states using equation (B 1).

(vi) Propose a new β' using a simple random walk, the step size $\sim N(0, 10^{-4})$.

(vii) Accept β' using a Metropolis–Hastings step with the acceptance probability

$$a = \min\left\{1, \frac{\pi(\beta')\Pr(\boldsymbol{Y}, \boldsymbol{C}|\beta')q(\beta' \to \beta)}{\pi(\beta)\Pr(\boldsymbol{Y}, \boldsymbol{C}|\beta)q(\beta \to \beta')}\right\}, \quad \text{(C 1)}$$

where $q(\beta \to \beta')$ is the proposal probability for β' from β which is the normal density for β' with mean β and variance 10^{-4}.

(viii) Repeat for ν' and d'.

(ix) Update each hidden state using a Gibbs update, drawing from the distributions given by the conditional probability of the states, determined by neighbouring states and observations, as described below.

(x) Determine $f(\boldsymbol{\theta})$ for the particular sequence of hidden states and parameters β and ν using expression (A 3).

(xi) Iterate by returning to step (iv).

(xii) Burn-in using 50 000 iterations. Use the following 50 000 updates to estimate the posterior probability distribution (using the ergodic average) of the hidden states (C_1, ..., C_n) and $f(\boldsymbol{\theta})$.

(xiii) Repeat steps 2–12 to construct 10 such Markov chains each with different initial values. Convergence tests showed that 50 000 updates were sufficient to get precise estimates of the parameters ($\hat{R} = 1.02$ for estimates of logit (proportion)) (Gelman *et al.* 2004, ch. 11.6).

(xiv) Use $10 \times 50\,000$ updates to determine the posterior probability densities of the model parameters.

The Gibbs update involves determining the conditional probability of the hidden states (given everything else). The assumption that the hidden states

are a first order Markov process means that the conditional probability of the hidden states is based only on neighbouring states and the corresponding datum. The full conditional probability of the hidden state, $C_n(k=2, \ldots, n-1)$, is given by

$$\Pr(C_k = i|C_{\setminus k}, \boldsymbol{y}) \propto \Pr(C_{k+1} = j|C_k = i) \times \Pr(C_k = i|C_{k-1} = h)\Pr(Y_k|C_k = i), \quad \text{(C2)}$$

where $C_{\setminus k}$ is the set of all states other than C_k; and i is the proposed value of the kth hidden state; and h and j are the current values of the hidden states $k-1$ and $k+1$, respectively.

The first and last states depend only on a single neighbour and the data associated with that state. That is

$$\Pr(C_1 = i|C_{\setminus 1}, \boldsymbol{Y}) \propto \Pr(C_2 = j|C_1 = i)\Pr(Y_1|C_1 = i), \quad \text{(C3)}$$

and

$$\Pr(C_n = i|C_{\setminus n}, \boldsymbol{Y}) \propto \Pr(C_n = i|C_{n-1} = h)\Pr(Y_n|C_n = i). \quad \text{(C4)}$$

The conditional probability of the states can be determined and this becomes the sampling distribution for the hidden state. Each of the n states can be updated in a forward, backward or random manner. To estimate values of ν and β, we do not need to infer hidden states. The simplified MCMC algorithm has the following steps:

(i) Assign the prior probability for β and ν using ($U[0, 0.1]$).
(ii) Initialize β, ν and d.
(iii) Determine the likelihood of the data using Baum's recursion formula.
(iv) Propose a new β' using a simple random walk, the step size $\sim N(0, 0.0001)$.
(v) Accept β' using a Metropolis–Hastings step with the acceptance probability

$$a = \min\left\{1, \frac{\pi(\beta')l(\boldsymbol{Y}|\beta')q(\beta' \to \beta)}{\pi(\beta)l(\boldsymbol{Y}|\beta)q(\beta \to \beta')}\right\}. \quad \text{(C5)}$$

(vi) Repeat for ν and d.
(vii) Iterate as above.

REFERENCES

Bailey, N. 1975 *The biomathematics of malaria*. London, UK: Charles Griffin.

Bartley, P. B., Schooneveldt, J. M., Looke, D. F., Morton, A., Johnson, D. W. & Nimmo, G. R. 2001 The relationship of a clonal outbreak of *Enterococcus faecium* VanA to methicillin-resistant *Staphylococcus aureus* incidence in an Australian hospital. *J. Hosp. Infect.* **48**, 43–54. (doi:10.1053/jhin.2000.0915)

Baum, L., Petrie, T., Soules, G. & Weiss, N. 1970 A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **41**, 164–171.

Boyce, J. M. 2001 MRSA patients: proven methods to treat colonization and infection. *J. Hosp. Infect.* **48**(Suppl. A), S9–S14. (doi:10.1016/S0195-6701(01)90005-2)

Bradley, S. J., Kaufmann, M. E., Happy, C., Ghori, S., Wilson, A. L. & Scott, G. M. 2002 The epidemiology of glycopeptide-resistant enterococci on a haematology unit: analysis by pulsed-field gel electrophoresis. *Epidemiol. Infect.* **129**, 57–64. (doi:10.1017/S0950268802007033)

Cooper, B. & Lipsitch, M. 2004 The analysis of hospital infection data using Hidden Markov models. *Biostatistics* **5**, 223–237. (doi:10.1093/biostatistics/5.2.223)

Cox, D. R. & Miller, H. D. 1965 *The theory of stochastic processes*. London, UK: Methuen.

D'Agata, E. M., Gautam, S., Green, W. K. & Tang, Y. W. 2002 High rate of false-negative results of the rectal swab culture method in detection of gastrointestinal colonization with vancomycin-resistant enterococci. *Clin. Infect. Dis.* **34**, 167–172. (doi:10.1086/338234)

Donskey, C. J., Hoyen, C. K., Das, S. M., Helfand, M. S. & Hecker, M. T. 2002 Recurrence of vancomycin-resistant enterococcus stool colonization during antibiotic therapy. *Infect. Control Hosp. Epidemiol.* **23**, 436–440. (doi:10.1086/502081)

Forrester, M. & Pettitt, A. N. 2005 Use of stochastic epidemic modeling to quantify transmission rates of colonization with methicillin-resistant *Staphylococcus aureus* in an intensive care unit. *Infect. Control Hosp. Epidemiol.* **26**, 598–606. (doi:10.1086/502588)

Gelman, A., Carlin, J., Stern, H. & Rubin, D. B. 2004 *Bayesian data analysis* 2nd edn., p. Fla. Boca Raton, FL: Chapman and Hall/CRC.

Giesecke, J. 1994 *Modern infectious disease epidemiology*. London, UK: Edward Arnold.

Gilks, W., Richardson, S. & Spiegelhalter, D. 1996 *Markov Chain Monte Carlo in practice*. London, UK: Chapman and Hall.

Green, P. 1995 Reversible jump Markov Chain Monte Carlo computation and Bayesian Model determination. *Biometrika* **82**, 711–732. (doi:10.1093/biomet/82.4.711)

Le Strat, Y. & Carrat, F. 1999 Monitoring epidemiologic surveillance data using hidden Markov models. *Stat. Med.* **18**, 3463–3478. (doi:10.1002/(SICI)1097-0258(19991230)18:24<3463::AID-SIM409>3.0.CO;2-I)

Lemmen, S. W., Hafner, H., Zolldann, D., Amedick, G. & Lutticken, R. 2001 Comparison of two sampling methods for the detection of Gram-positive and Gram-negative bacteria in the environment: moistened swabs versus rodac plates. *Int. J. Hyg. Environ. Health* **203**, 245–248. (doi:10.1078/S1438-4639(04)70035-8)

Lodise, T. P., McKinnon, P. S., Tam, V. H. & Rybak, M. J. 2002 Clinical outcomes for patients with bacteremia caused by vancomycin-resistant enterococcus in a level 1 trauma center. *Clin. Infect. Dis.* **34**, 922–929. (doi:10.1086/339211)

MacDonald, I. & Zucchini, W. 1997 *Hidden Markov models for discrete valued time series*. London, UK: Chapman and Hall.

Murray, B. 2006 Overview of enterococci. In *UpToDate* (ed. B. D. Rose). Massachusetts, MA: UpToDate. (CD-rom.)

Pelupessy, I., Bonten, M. J. & Diekmann, O. 2002 How to assess the relative importance of different colonization routes of pathogens within hospital settings. *Proc. Natl Acad. Sci. USA* **99**, 5601–5605. (doi:10.1073/pnas.082412899)

Petrushin, V. 2000 Hidden Markov models: fundamentals and applications. Part2 Discrete and continuous hidden Markov models. In *Online Symp. for Electronics Engineers*. http://www.techonline.com/osee/.

Reisner, B. S., Shaw, S., Huber, M. E., Woodmansee, C. E., Costa, S., Falk, P. S. & Mayhall, C. G. 2000 Comparison of three methods to recover vancomycin-resistant enterococci (VRE) from perianal and environmental samples collected during a hospital outbreak of VRE. *Infect. Control Hosp. Epidemiol.* **21**, 775–779. (doi:10.1086/501734)

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. (doi:10.1214/aos/1016120364)

Suppola, J. P., Kolho, E., Salmenlinna, S., Tarkka, E., Vuopio-Varkila, J. & Vaara, M. 1999 vanA and vanB incorporate

into an endemic ampicillin-resistant vancomycin-sensitive *Enterococcus faecium* strain: effect on interpretation of clonality. *J. Clin. Microbiol.* **37**, 3934–3939.

Tenover, F., Arbeit, R., Goering, R., Mickelsen, P., Murray, B., Persing, D. & Swaminathan, B. 1995 Interpreting chromosomal DNA restriction patterns produced by pulsed field gel electrophoresis: criteria for bacterial strain typing. *J. Clin. Microbiol.* **33**, 2233–2239.

Trick, W. E., Kuehnert, M. J., Quirk, S. B., Arduino, M. J., Aguero, S. M., Carson, L. A., Hill, B. C., Banerjee, S. N. & Jarvis, W. R. 1999 Regional dissemination of vancomycin-resistant enterococci resulting from interfacility transfer of colonized patients. *J. Infect. Dis.* **180**, 391–396. (doi:10.1086/314898)

Trick, W. E., Paule, S. M., Cunningham, S., Cordell, R. L., Lankford, M., Stosor, V., Solomon, S. L. & Peterson, L. R. 2004 Detection of vancomycin-resistant enterococci before and after antimicrobial therapy: use of conventional culture and polymerase chain reaction. *Clin. Infect. Dis.* **38**, 780–786. (doi:10.1086/381552)

Weinstein, J. 2005 Hospital-acquired (nosocomial) infections with vancomycin-resistant enterococci. In *UpToDate* (ed. B. D. Rose), p. 2006. Massachusetts, MA: UpToDate. (CD-rom.)