List of Tables

List of Figures

Chapter 8

Longitudinal models of ESBL-E carriage

8.1 Chapter Overview

In this chapter, I develop time-inhomogeneous Markov models to model ESBL-E carriage and fit them using the Bayesian probabilistic programming language Stan. I demonstrate that fitting these models is feasible with modest computational requirements, and that they are very flexible. I find that hospitalisation acts to increase both rate of ESBL-E acquisition and loss, with a net effect of rapidly increasing ESBL-E carriage prevalence and that antibacterial exposure acts to prolong ESBL-E carriage by reducing the rate of ESBL-E loss. However, it is the synergistic effect of hospitalisation and antibacterial exposure that seems to drive the rapid increase in ESBL-E carriage prevalence observed in antibacterial-exposed inpatients; I also find that co-trimoxazole preventative therapy (CPT) likely plays an important role as a determinant of long-term ESBL-E carriage. The models I develop also support a post-exposure effect of antibacterials, such that they continue to have an effect long after they would be expected to be excreted from the body. I present hypotheses about the mechanism of such an effect along with the implications of my findings for antimicrobial stewardship interventions, and planned further work.

8.2 Introduction and chapter aims

In Chapter 5, I presented the longitudinal ESBL-E carriage data for the three arms of the clinical study that underpin this thesis. Antibacterial-exposed, hospitalised participants (arm

1) showed a rapid increase in ESBL-E carriage prevalence, whilst antibacterial-unexposed hospitalised participants (arm 2) showed a much more modest increase. This suggests that antimicrobial exposure is the most significant determinant of acquisition of carriage; however, this unadjusted analysis is open to confounding. The participants recruited to the two arms of the study differ in important characteristics: antimicrobial unexposed participants are younger, less likely to be HIV-infected, with less cotrimoxazole preventative therapy (CPT) exposure, and crucially, a shorter length of hospital stay. An attempt to adjust for potential confounders using simple logistic regression models failed; in this chapter I develop longitudinal models to quantify the relative roles of antibacterial exposure and hospitalisation in driving ESBL-E carriage.

There have been few prior attempts to model longitudinal ESBL-E carriage, and none where the focus was on he role of antimicrobials. Three attempts, all from the Netherlands, have taken a variety of approaches: by fitting a Weibull distribution to community sample data[1], by fitting a beta-distribution of admission and discharge carriage probability to data from trials of contact precautions in Dutch hospitals[2], and by modelling household ESBL-E acquisition as a Markov process[3]. None of these studies included the effect of antibacterial exposure as a covariate. The Markov model approach is an attractive method to model multi state interval censored data[4], has been used with a variety of clinical datasets[[5]; Andersen1988; Jackson2002] and is implemented in the *msm* package in R[6] where a maximum likelihood method is used to fit the models. However, *msm* allows only stepwise-constant covariate effects, largely for reasons of computational tractability in the maximum-likelihood framework; there is no reason to assume that the effect of, say, antibacterial exposure will act in this fashion. The aims of this chapter, therefore, are:

- To generalise *msm*-type models to allow true time-varying covariates.
- To demonstrate the feasibility of fitting such models.
- To used the fitted models to infer an unbiased estimate of the relative roles of antibacterial exposure and hospitalisation in driving ESBL-E carriage by both fitted parameter estimates and simulating different levels of exposure.
- To compare models with and without a post-exposure effect of antibacterials to assess the support in this data for such an effect
- To combine the models with ESBL-E species data and with the WGS isolate typing presented in Chapter 7 to explore carriage at the level of species, *E. coli* clone and ESBL-containing mobile genetic element.

8.3 Methods

8.3.1 Developing the models used in this chapter

In the broadest sense when constructing a model, our aim is to estimate the most likely values of the parameters of the model, θ, given the data we have, x. The starting point for estimating likely parameter values, given a choice of model, is usually the *likelihood*: this is the probability of the data, given a set of parameter values. In standard probability notation, this is written as $P(x|\theta)$. In fact, this is not the quantity we are interested in; we would like to know $P(\theta|x)$: the probability of the parameter values, given the data. Both frequentist and Bayesian modelling approaches provide methods to estimate this quantity, but the starting point for both is the likelihood, $P(x|\theta)$, because it is usually much more straightforward to derive an expression for $P(x|\theta)$ rather than $P(\theta|x)$. I will here derive a general likelihood for a two state intermittently observed process; in order to use this likelihood, it is necessary to make some assumptions about the data generating process. I have chosen to use a Markov model, and I will then derive the likelihood for this model, describe how covariates will be incorporated , describe how the model was fit - the process taking us from the likelihood to the most likely parameter values - and finally how goodness of was fit assessed.

8.3.2 General form of likelihood

First, I derive a general expression for the likelihood of a two-state intermittently observed process without making any assumptions about the model structure or functional form. Assume we have N participants with any given participant n in a state $S_n(t)$ at time t: either ESBL-E colonised ($S_n(t) = 1$) or uncolonised ($S_n(t) = 1$). For each participant n we have a number of measurements of $S_n(t)$ at a number of time points. The number of measurements varies for each participant, and can be denoted by j_n, making the time of measurements $t^n_{j_n}$ for participant n; and so for each participant we know the j_n values $S_n(t^n_{j_n})$.

To arrive at the likelihood for these observations, consider first the simplest situation that we have: the measurements of ESBL status at two time points, t_A and t_B for a single participant, n. The likelihood we wish to calculate, in words, is the probability of the participant being in the second observed state at time t_B, given they were in the first state at t_A and given the parameters of the model, θ. Or, mathematically:

$$P(S_n(t_B)|S_n(t_A), \theta) \tag{8.1}$$

Assuming all the observations are independent, the probability of all of the states we have

observed for this participant is the product of all the probabilities of the individual states:

$$\prod_{k=2}^{j_n} P(S_n(t_k^n)|S_n(t_{k-1}^n), \theta) \tag{8.2}$$

And the probability of observing the data we have is then simply the product of the probability of all the individual transitions:

$$\prod_{n=1}^{N} \prod_{k=2}^{j_n} P(S_n(t_k^n)|S_n(t_{k-1}^n), \theta) \tag{8.3}$$

This is the quantity that we wish to calculate: the likelihood for the observed data, $P(x|\theta)$. Note that the sum over states for an individual in equation (8.3) starts from 2; if a participant has only one available sample then this does not provide any information about transition probabilities, and must be excluded from the analysis.

8.3.3 Markov model likelihood

In order to calculate the likelihood, we need to make some assumptions about the data generating process. In this case, I have chosen to use a Markov model. Markov models are defined by instantaneous transition probabilities, analogous to the hazard of death in a survival model, which is a simple two-state Markov system. Unlike a survival model (where it is not possible to move from the death state to alive), a general Markov model is defined by a transition hazard from each state to each other state in the system. These are traditionally expressed as a Q matrix of instantaneous transition intensities[Hout2016; Jackson2011a] (assuming a two-state system):

$$\mathbf{Q}(t) = \begin{pmatrix} q_{00}(t) & q_{01}(t) \\ q_{10}(t) & q_{11}(t) \end{pmatrix} \tag{8.4}$$

Where q_{ij} represents the instantaneous transition intensity from state i to state j. The rows of the Q-matrix must sum to 1 (every participant has to be in one state or another), so if we define the hazard of ESBL-E acquisition to be λ and the hazard of ESBL-E loss to be μ (8.1), the Q-matrix becomes, in our case:

$$\mathbf{Q}(t) = \begin{pmatrix} -\lambda(t) & \lambda(t) \\ \mu(t) & -\mu(t) \end{pmatrix} \tag{8.5}$$

However, we are not interested in the Q-matrix *per se* but rather the probability p_{ij} of starting in state i at time 0 and being in state j at time t; this can be written in matrix notation as $\mathbf{P}(t)$ and is related to $\mathbf{Q}(t)$ by the differential equations:

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{Q}(t).\mathbf{P}(t) \tag{8.6}$$

Where $\mathbf{Q}(t).\mathbf{P}(t)$ is the matrix product of $\mathbf{Q}(t)$ and $\mathbf{P}(t)$. In order to eva,uate $\mathbf{P}(t)$, therefore we need to solve this system of differential equations. However, there are limited situations in which these equations have analytic solutions. If the system has time constant or piecewise constant Q matrix the matrix exponential is a solution:

$$\mathbf{P}(t) = e^{\mathbf{Q}} \tag{8.7}$$

However, there is no reason to suspect particularly that the effect of covariates on ESBL-E carriage (e.g. antimicrobials) would be stepwise constant and so a more flexible model is needed. For general time-varying transition intensities, there is no analytic solution to the above equations. However, all is not lost: we can express the likelihood in terms of the differential equations defined by the equations above and solve them numerically in order to calculate the likelihood. The matrix notation above can be simplified, assuming that the system starts in state 1 or 0:

$$\frac{dP_0(t)}{dt} = -\lambda(t)P_0(t) + \mu(t)P_1(t) \tag{8.8}$$

$$\frac{dP_1(t)}{dt} = \lambda(t)P_0(t) - \mu(t)P_1(t) \tag{8.9}$$

Where $P_i(t)$ is the probability of being in state i at time t. Numerical ordinary differential equation (ODE) solvers can quickly solve these equations to calculate, for example, $P(S_n(t_B)|S_n(t_A), \theta)$ from the simplest example above: the probability that a participant n at time t_B is in a given state, given that they were in state $S_n(t_A)$ at time t_A, and given the parameters θ. This calculation can be completed for all measurements and participants, resulting in the likelihood of the system, $P(x|\theta)$.

In order to use this model for inference, two questions must be addressed: first, how to incorporate time-varying covariates; and second, how to practically fit the model. I address each of these questions below.

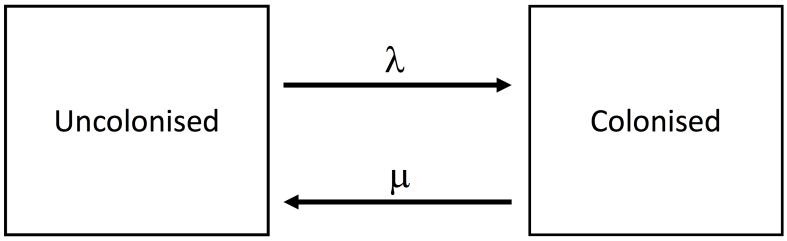

Figure 8.1: Two state ESBL-E model showing instanteneous hazard of ESBL-E acquisition (λ) or loss (μ).

8.3.4 Incorporating covariates: a proportional hazard model

I have chosen to incorporate covariates using a proportional-hazards model, following both Marshall and Jones[7] and the *msm* package in R[6]. In this model the transmission intensities become:

$$\lambda(t) = \lambda_0 \exp\left(\beta_0 x_0(t) + \beta_1 x_1(t) + ...\beta_m x_m(t)\right) \tag{8.10}$$

$$\mu(t) = \mu_0 \exp\left(\alpha_0 x_0(t) + \alpha_1 x_1(t) + ...\alpha_m x_m(t)\right) \tag{8.11}$$

Where the $x_k, k = 1, 2...m$ are the m time-varying covariates in the model and the coefficients α_k and β_k are the coefficients of these covariates; these have a straightforward interpretation in that the exponential, e^{α_k} ore^{β_k} can be interpreted as a hazard ratio, as per a simple survival model.

An assumption then needs to be made about the functional form of x_m. In a stepwise-constant covariate model in which an exposure occurs between t_A and t_B, $x(t)$ would take the value 1 for all $t_A \leq t \leq t_B$t and 0 at other times, meaning that the effect of the exposure does not persist once it ceases. Though this may be plausible for some exposures, it seems possible that antimicrobial exposure (for example) might have a longer lasting effect or post-exposure effect; in order to explore this possibility, it is necessary to decide on a flexible, plausible, functional form that such an effect might take. I have decided to use an exponential function, such that:

$$x_k(t) = \begin{cases} 0 & \text{if } t < t_A \\ 1 & \text{if } t_A \leq t \leq t_B \\ \exp\dfrac{-(t - t_B)}{\gamma_k} & \text{if } t > t_B \end{cases} \tag{8.12}$$

Where the parameter γ_k is a model parameter for each of the covariates, to be estimated from the data, and is related to the half life, $t^k_{\frac{1}{2}}$ of the decay of the effect of the exposure by:

$$t^k_{\frac{1}{2}} = \gamma_k \ln(2) \approx 0.69\gamma_k \tag{8.13}$$

From the definition of the half life of an exponential decay process. This parametrisation has the advantage that the data can fit the size of the parameters γ_k; if the data are more in keeping with a stepwise effect of the covariates, then a small ($\ll 1$) γ would approximate a step function and this could be fit by the model. Alternatively a larger would result in the effect of the covariate persisting after exposure, but decaying over time. This allows us to test the hypothesis that antimicrobial exposure (for example) has an effect that persists once exposure finishes, by both the magnitude of the fitted γ_k, and comparing stepwise-constant covariate models to models incorporating the γ_k parameters.

The parameters of the model all have the advantage of having a reasonably intuitive meaning: $\exp(\alpha)$ and $\exp(\beta)$ are the hazard ratio for ESBL-E loss and acquisition, respectively; the reciprocals of λ and μ are the mean time in days spent in the uncolonised or colonised states, respectively; and $\ln(2)\gamma \approx 0.69\gamma$, as stated above, is the half life of the post-exposure effect.

8.3.5 Building and fitting models

The Bayesian probabilistic programming language *Stan* incorporates an ordinary differential equation solver, and will allow the fitting of the model in a Bayesian framework[8]. In this framework, Bayes' rule allows us to estimate our probability distribution of interest, $P(\theta|x)$, called the *posterior* in the Bayesian framework, a long as we provide a *prior*, encoding our prior beliefs about the values of the parameters as a probability distribution for each parameter[9]. Stan then uses the No-U-Turn Sampler(NUTS) implementation of Markov-chain Monte-Carlo (MCMC) sampling[10] to sample from the posterior to provide $P(\theta|x)$. It can be shown that, given infinite chain length, MCMC estimates are guaranteed to be unbiased samples from the posterior; when this occurs the chains have said to converged. Unfortunately there is no diagnostic test that guarantees convergence, rather tests that are necessary but not sufficient to ensure convergence: running multiple chains from different starting points with examination of traceplots to show within and between mixing of chains, and the $\hat{R}$ statistic, which measures mixing of the two halves of an MCMC chain. At convergence, $\hat{R}$ should be close to 1[9]. In addition, divergences - failure in the NUTS sampler - can be indicative of difficult topography in the posterior at the area where the divergences occur, and suggest that parameter estimates may be biased, and are flagged by Stan. All of these tests were used to diagnose convergence.

Two decisions must be made in order to fit the model: covariates must be chosen to include and priors specified. Models were built sequentially to predict ESBL-E status, starting from the simplest possible, then adding complexity:

- *Model 1:* Composite antibacterial variable (includes all antibacterials) and hospitalisation variable as explanatory variables, both included with stepwise constant effect and no post exposure effect.
- *Model 2:* As per model 1 except antibacterial exposure modelled with decaying post-exposure effect.
- *Model 3:* Hospitalisation, TB therapy and co-trimoxazole exposure all modelled as stepwise constant covariates. All other antibacterials included in a composite variable with decaying post-exposure effect.
- *Model 4:* Hospitalisation, TB therapy and co-trimoxazole exposure all modelled as stepwise constant covariates; ceftriaxone, ciprofloxacin and amoxicillin exposure included in a composite variable with decaying post-exposure effect, with γ allowed to vary for each agent.

Weakly informative priors were used. A normally distributed prior centred at 0 with standard deviation 2 was used for all the α and β parameters. A parameter value of 2 corresponds to a hazard ratio of 7.4; it would be surprising if any effect is greater than this so this could be argued to be a weakly informative prior. Normally distributed priors centred at 0 with standard deviation 0.2 were used for the μ and λ parameters; in a model with no covariates, the inverse of these parameters are the mean times that an individual would remain in the colonised or uncolonised states, respectively, so a value of 0.2 corresponds to a mean state occupancy time of 50 days. A normally distributed prior centred at 0 and with standard deviation 50 days was used for all γ parameters.

The Stan code for the models is given in the appendix to this chapter. Four chains were run in each case, with a warmup of 500 iterations and run for 1000 iterations in total. Convergence was assessed using the diagnostics described above. Stan v2.19 was used to sample from the posterior, accessed via Rstan v2.19.2, and run on the Wellcome Sanger Institute computing cluster under Linux Red Hat v7.6, running R v3.5.3, and gcc v6.3.0 C++ compiler. Four cores and 3GB of memory per model fit were used. Posterior samples were brought to my local machine (MacBook Pro running mac OS Mojave 10.14.5) and further analyses undertaken with R3.6.0.

8.3.6 Assessing goodness of fit

Model goodness of fit was assessed in two ways; first, by graphical posterior predictive checks: comparing predicted total number of ESBL-E positive samples to the actual number across the three arms. This was done by using the posterior parameter estimates for each MCMC draw (after discarding warmup samples) to generate a predicted probability of the ESBL-E positive state for each data point, then sampling from a Bernoulli distribution to convert to predicted state occupancy. Each data point therefore had 2000 predictions for state occupancy, one for each posterior draw. These were plotted as kernel density plots against actual state occupancy, stratified by arm, to visualise the goodness of fit of the model, and to compare between models.

Second, models were compared using leave-one-out cross validation, as implemented in the *loo* v2.1.0 package in R[11]. This estimates the out-of sample predictive ability of the model by estimating a quantity called the expected log pointwise predictive density ($ELPD$) essentially the log of the likelihood for a new, unseen dataset conditional on the current data. This quantity is estimated using leave-one-out cross validation to produce and estimate of the $ELPD$, hereafter referred to as $ELPD_{loo}$. The standard error of $ELPD_{loo}$ for a model is also calculated and so two models can be compared by comparing the $ELPD_{loo}$ difference and standard error; if the difference is greater than twice the standard error (i.e. a 95% confidence interval, assuming normality) we can be confident that one model would be expected to have greater out-of-sample predictive ability than the other[11]. Because this technique estimates out-of-sample predictive ability it naturally incorporates a penalty for including multiple parameters and hence overfitting, as an overfit model would be expected to have worse out of sample predictive ability and hence lower $ELPD_{loo}$.

8.3.7 Exploring differences in carriage dynamics by bacterial species and *E. coli* genotype

The models fit as described above predict whether a participant will be colonised with any ESBL producing organism at a given time point, but this classification obscures a lot of complexity. A participant can be colonised with different ESBL- producing species (largely *Eschericha coli* or *Klebsiella pneumoniae*), and different clones of those species containing different ESBL genes on different mobile genetic elements (MGEs). It may be that there is heterogeneity in carriage dynamics across these different levels of the system. To address this hypothesis, the best fitting model identified from the four described above was refit but the "colonised" state modified to either consider the species level or to use the whole genome sequence data presented in Chapters 6 and 7 as a high-resolution typing system

to track bacteria through he system. The analysis in Chapter 7 suggests that the element most conserved within participants is the bacterial clone-ESBL contig combination, where the bacterial clone clusters were defined with the hierarchical BAPS algorithm and the ESBL-contig clusters defined with the cd-hit algorithm, as described in Chapter 7. The hierBAPS cluster-contig cluster pairs are coded as follows in this chapter: a.ESBLgene.b where a is the ID number of the level 2 hierBAPS cluster, and ESBLgene.b is the number of the contig cluster for a given ESBL gene, and for the rest of the chapter for brevity each unique hierBAPS cluster-contig cluster will be referred to as an *E. coli* genotype. All *E. coli* genotypes which were identified in more than 15 samples - 6 in total - were included and so the models were refit defining the colonised state as the presence of, respectively, ESBL *E.coli* , *K. pneumoniae* or one of the six included hierBAPS cluster-contig cluster pairs. The parameters for these models were compared with each other and with the original ESBL model.

8.3.8 Simulations from the posterior

Finally, in order to better understand the relative role of antimicrobial exposure and hospitalisation in driving ESBL-E carriage, I conducted simulations with these exposures set at varying levels. The probability of ESBL colonisation as a function of time was calculated by solving the equations (8.8) and (8.9) using the R package *deSolve* v1.2.4[12], for each of the 2000 posterior parameter estimates from the posterior and assuming a 50% initial probability of ESBL colonisation. This yielded a distribution of carriage probability at each time point which was summarised using the median and 95% confidence intervals and plotted against time for varying covariate values: days of hospitalisation was varied from one to twenty in steps of five, as was antimicrobial exposure, and each simulation repeated both with and without CPT.

8.4 Results

8.4.1 The effect of antibacterials and hospitalisation on ESBL-E carriage

First, I fit the four models with three aims: to identify the model that provides the best trade off between predictive ability and the computational cost to fit; to explore the relative effects of hospitalisation versus antimicrobial exposure on ESBL-E carriage by assessing the posterior parameter values of these models; and to assess support in the data for a post-antibacterial effect on ESBL-E carriage that persists once antibacterial therapy is stopped. For these models, the colonised state was defined as at least one ESBL producing organism of any species identified in a sample, and uncolonised as no ESBL producer identified. After

excluding participants with only one sample, there were 993 pairs of samples in 363 participants remaining that contributed data to the analysis. All four models converged within the 1000 iterations; $\hat{R}$ was less than 1.1 for all parameters and all traceplots showed good mixing of chains. There were no divergences of the NUTS MCMC sampler in any of the models. There was a computational cost to increasing the number of parameters, as would be expected from the increase in dimensionality of the posterior: model one took 3.5 hours to fit, model two 13.7 hours, model three 17.1 hours and model four 33.4 hours.

The parameter estimates for the models are shown in Figure 8.2. There were significant correlations between some posterior parameters (see Figure 8.6 in the Chapter Appendix for pairwise plots for model 2 as an example): particularly λ and μ, and the α and β parameters. This is not necessarily problematic in that it is not necessarily a source of bias, but can make it difficult for some MCMC algorithms (e.g. Metropolis-Hastings) to adequately sample from the posterior[Gelman]. Nevertheless, the diagnostics suggest that the Stan NUTS sampler had no problems.

The effect of hospitalisation is consistent across all models; in most models, the 95% credible intervals for both α and β for hospitalisation do not cross zero and are positive, suggesting that the hazard ratio of hospitalisation on both the rate of acquisition and loss of ESBL-E is very likely to be greater than one, and the effect of hospitalisation is to increase both the rate of acquisition and loss of ESBL-E. The estimated effect sizes are consistent across the models though, as expected, uncertainty in the estimate increases as more parameters are added to the model.

The effect of antibacterial exposure is also reasonably consistent across the models; the parameter α is negative in all cases, and often the 95% credible intervals do not cross zero, suggesting that the hazard ratio of antimicrobial exposure is likely to be less than zero. The effect sizes are similar in all cases, for all agents (including TB therapy), whether antibacterial exposure is considered as an aggregate variable or as individual agents; though in the extreme case where agents are all considered individually (Model 4, Figure 8.2D) the uncertainty in the estimates makes it difficult to draw any firm conclusions. This suggests that all the considered antibacterial agents act, with broadly similar effect size, to prolong ESBL-E carriage by reducing the rate of loss. No β parameter (the log hazard ratio of ESBL-E acquisition) has 95% credible intervals that do not cross zero, consistent with antibacterial exposure have no or limited effect on ESBL-E acquisition.

The relative predictive ability of the four models were assessed in two ways: first, the predicted proportion of ESBL-E positive samples were plotted by sampling from the posterior (Figure 8.3); second, the pairwise $ELPD_{loo}$ differences (and standard errors in the differences) between all models calculated (Table 8.1). All models predicted ESBL-E carriage reasonably poorly for

Table 8.1: Estimates (and standard error) of pairwise expected log pointwise predictive density differences for all models

	Model 1	Model 2	Model 3	Model 4
Model 1	0.0 (0.0)	10.5 (4.2)	10.0 (6.4)	15.0 (7.0)
Model 2	-	0.0 (0.0)	-0.5 (5.2)	4.4 (6.0)
Model 3	-	-	0.0 (0.0)	4.9 (3.7)
Model 4	-	-	-	0.0 (0.0)

Note:
Cells in table compare row model to column model. A positive number favours the model in the column. The standard error of the ELPD difference is given in brackets; if twice the standard error is less than the estimated ELPD difference then we can be confident that the column model has better out-of-sample predictive fit than the row model. All models have better fit than model 1 but models 2-4 all have similar fit.

arm two and three participants, but better for arm one 8.3). The addition of a post-antibiotic effect improved model fit (seen by comparing model 1 to model 2) but models two, three and four, had similar fit despite the increase in number of parameters from seven in model two to seventeen parameters in model four. Model two therefore provides a good balance between computational tractability, interpretation and predictive ability; the parameter estimates for this model, expressed as hazard ratios for α and β, the mean time in state for λ and μ and half life of post-antibacterial effect for γ are shown in Table 8.2.

8.4.2 Exploring bacterial species and genotype differences in carriage dynamics

Next, I explored the differences in carriage dynamics between ESBL-E species and *E. coli* genotype, by refitting model 2 but considering the colonised/uncolonised states to be, in turn, presence or absence of *E. coli*, *K. pneumoniae* or any of the top six most prevalent *E. coli* genotypes (as defined by the combination of ESBL containing contig cluster and *E. coli* hierBAPS cluster [Chapter 7]), and refitting the model for each one. All 993 within-participant sample-pair comparisons were used to fit the *E. coli* and *K. pneumoniae* models, but because sample collection continued after the sequenced *E. coli* included here were shipped, all samples collected after this time were excluded. 585 samples from 251 participants were therefore included in the genotype models.

The parameter estimates for these eight models (alongside the original ESBL-E presence/absence model) are shown in figure 8.4. In general, there was more uncertainty in the

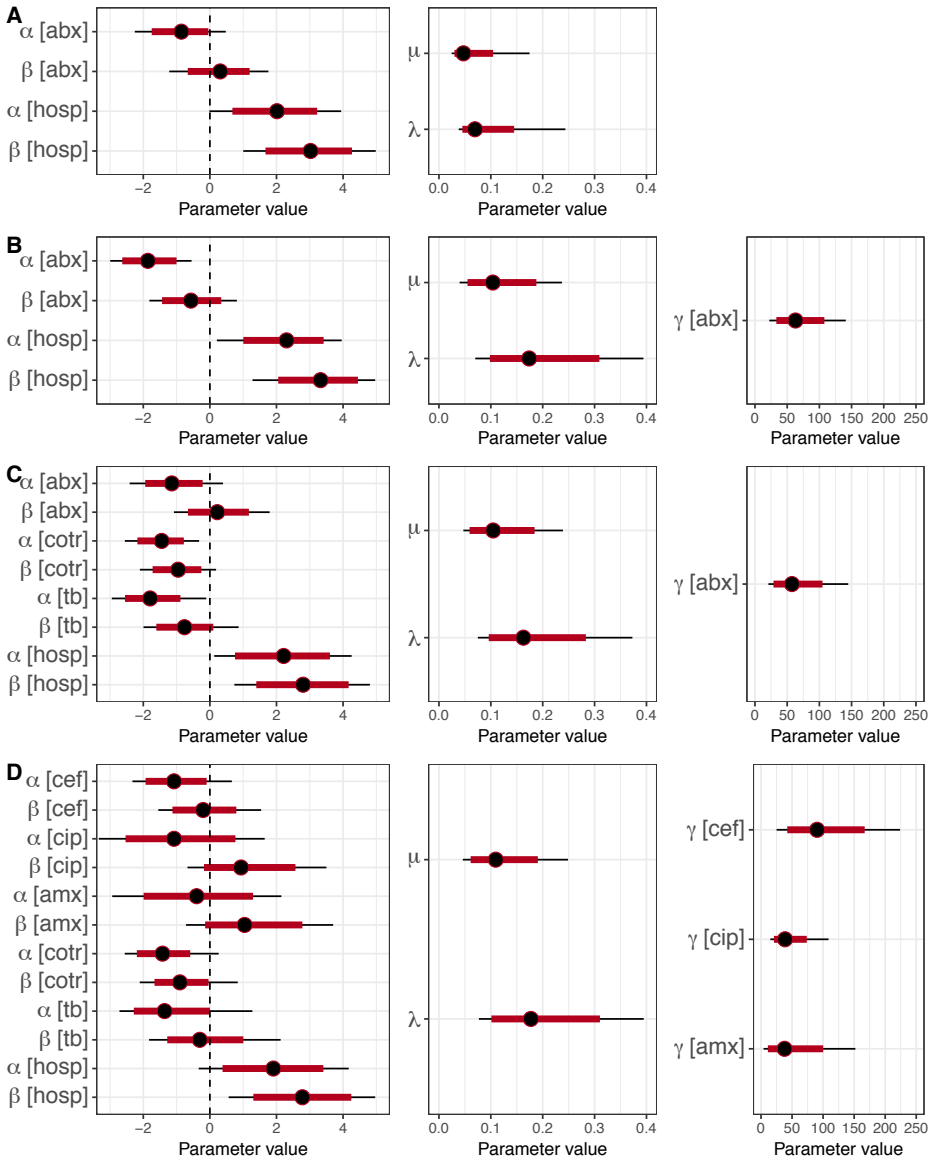

Figure 8.2: Parameter estimates from increasingly complex Markov models to predict ESBL carriage. Black lines are 95% and red lines 80% credible intervals. A: Model 1 includes stepwise constant covariates only, animicrobial exposure and hospitalisation. λ is the baseline hazard and β the log hazard ratio of ESBL-E acquisition, μ the baseline hazard and α the log hazard ratio of ESBL-E loss. B: Model 2 adds a post-exposure effect of antimicrobial exposure, parameterised by γ as described in the text. C: Model 3 adds stepwise constant covariates for TB therapy (tb) and cotrimoxazole (cotri) with all other antimicrobial exposure captured in the abx variable, which has a post exposure effect as before. D: Model 4 seperates the effect of antimicrobial exposure into the component agents, with post exposure effects for all except cotrimoxazole and TB therapy. In most models 95% credible intervals of α[hosp] and β[hosp] do not cross zero and are positive, suggesting that hospitaliation acts to both increase rate of ESBL-E acquisition and loss; for antimicrobial exposure, on the other hand, only the 95% for antimicrobial α values consistently do not cross zero, and are negative, suggesting that the effect of antimicrobial exposure is to reduce the rate of ESBL-E loss. It is also clear that adding parameters to the model increases the uncertainty in the estimates (e.g. compare model 2, B, to model 4, D).

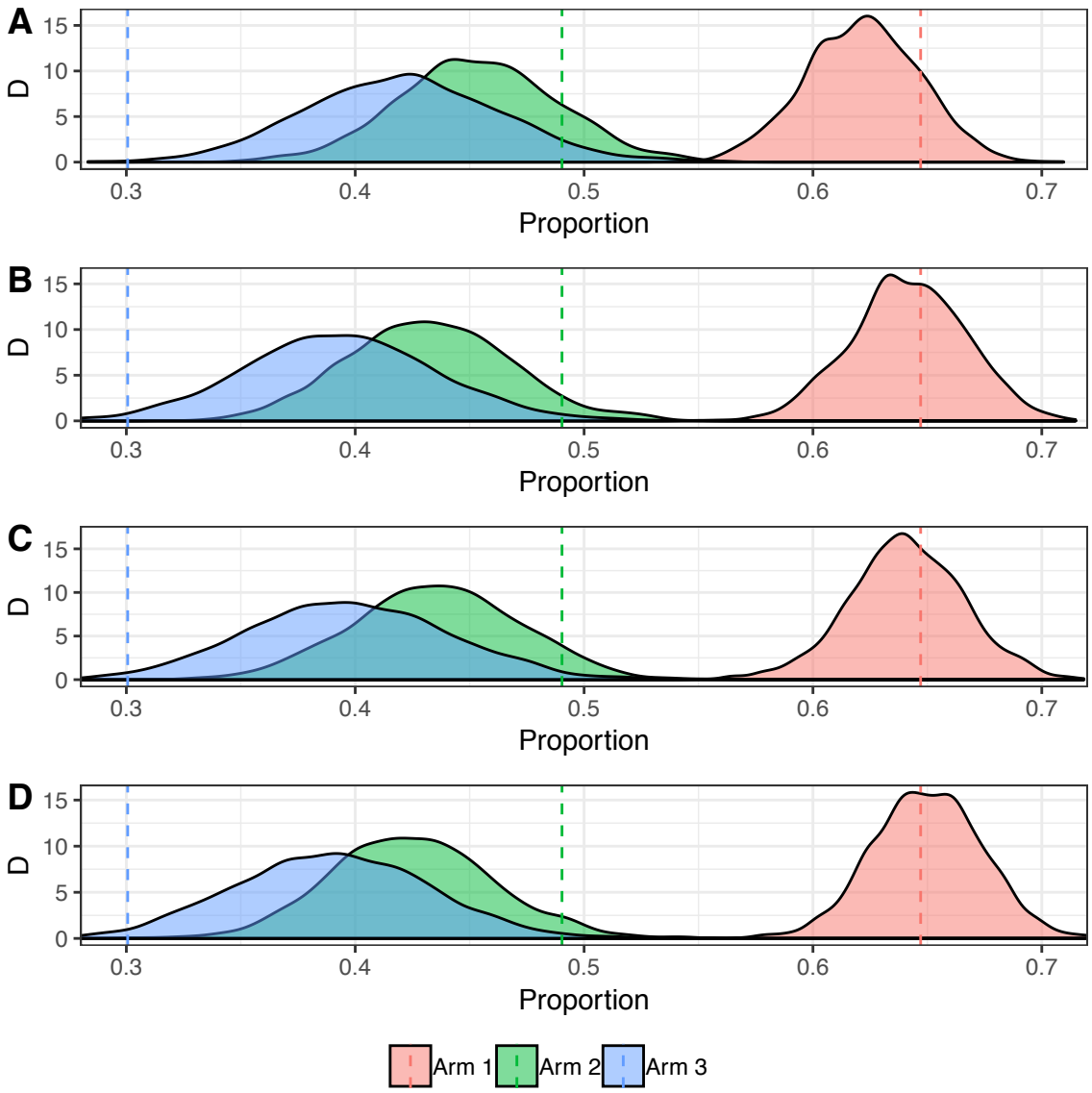

Figure 8.3: Posterior predictive checks: kernal density estimate, D, of predicted proprtion of ESBL-E positive samples, stratified by arm for Model 1 (A), Model 2 (B), Model 3 (C) and Model 4 (D), generated by sampling from a Bernoulli distribution using the predicted probability for each sample (n=993) for each draw from the posterior, excluding warmup draws (n = 2000). True proportion of ESBL-E positive samples are shown for each arm by dotted vertical line. In all cases, predictions are poor for arm 2 and 3 samples, but the addition of a post-antibacterial effect (quantified by γ) improves fit, especially in arm 1 participants: compare Model 1 (A) with stepwise constant covariates to Model 2 (B) with post-antibacterial effect. Models 2-3 (B-D) have similar predictions despite more parameters.

Table 8.2: Parameter estimates (and 95% confidence intervals) from model 2

Variable	Value
Effect of Antibacterials	
Hazard ratio for ESBL-E loss	0.16 (0.05-0.58)
Hazard ratio for ESBL-E acquisition	0.57 (0.16-2.25)
Effect of Hospitalisation	
Hazard ratio for ESBL-E loss	10.01 (1.24-52.34)
Hazard ratio for ESBL-E acquisition	27.82 (3.60-143.18)
Post Antibacterial Effect	
Half life (days)	43.67 (15.42-97.66)
Mean time in state	
Uncolonised (days)	9.65 (4.22-25.07)
Colonised (days)	5.76 (2.54-14.30)

Note:
Hazard ratios are the exponential of the parameters α and β in the model; half life is equal to log(2) multiplied by γ; mean time in state assumes all other covariates are equal to zero and is then the reciprocal of λ or μ.

parameter estimates for the new models, as might be expected as there are fewer carriage events, and fewer samples in the case of the genotype models. The only significant parameter difference between the models was in the λ parameter, the baseline hazard of state acquisition. The magnitude of the difference was large; for example the median (95% CI) λ_{ESBL} estimate of 0.10 (0.07-0.15) is almost three orders of magnitude larger than the estimate of $\lambda_{6.CTXM.27}$, 0.002 (0.001- 0.003). These values would correspond to a mean (95% CI) time in the uncolonised state of 10 (6-14) days for the ESBL model versus 500 (333-1000) days for the genotype model, assuming all other covariates were zero. The hazard rate of state loss, μ was similar, however, meaning that the time in the colonised state is similar for the ESBL model, and for all the *E. coli* genotype models.

8.4.3 Simulation of different antibacterial and hospitalisation scenarios

Finally, to better understand the relative roles of antimicrobial exposure and hospitalisation in driving ESBL-E carriage, I simulated the probability of ESBL-E colonisation as antibacterial and hospital exposure changed from 1 to 20 days, assuming a 50% baseline probability of ESBL-E colonisation (Figure 8.5) and both with and without cotrimoxazole preventative therapy. Model 2 was used for these simulations. Hospitalisation seems to rapidly increase in carriage probability and antimicrobial exposure produces a slower rise. Most striking, however, is that the effect of both exposures simultaneously causes a rapid increase in ESBL-E

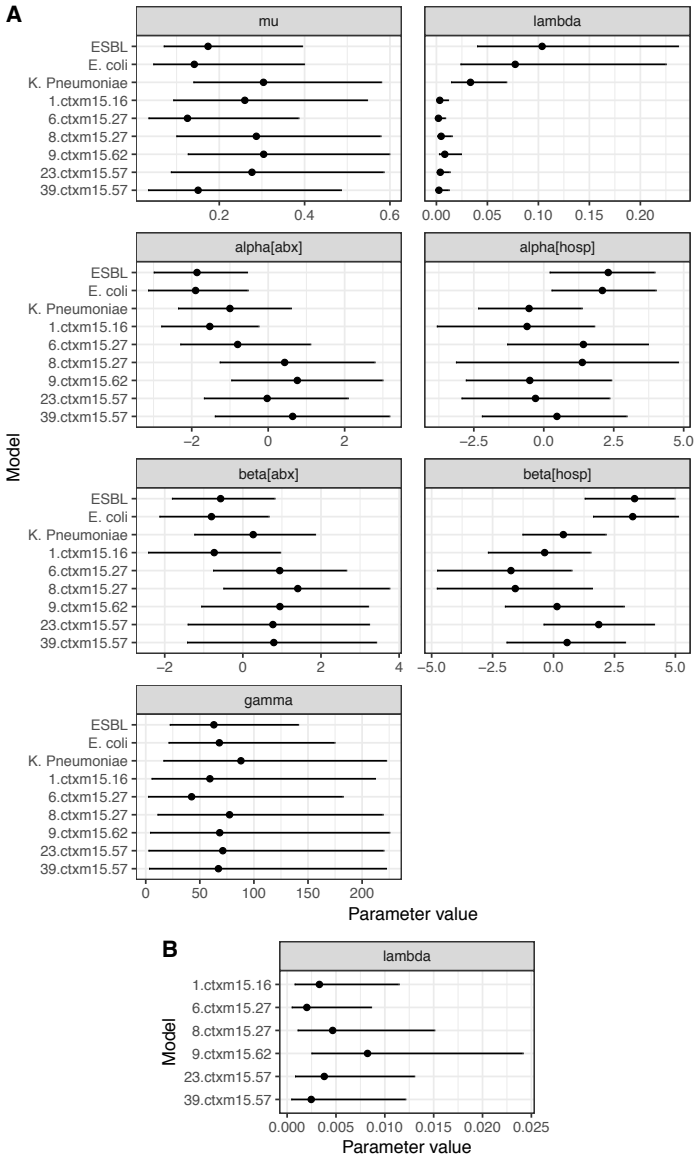

Figure 8.4: Parameter estimates from two state models predicting species and E. coli genotype carriage, compared to original model, which predicted carriahe of any ESBL-E. A: All parameters, showing that the only significant difference between the models is the parameter λ (the hazard of acquisition), with an order of magnitude difference between the hazard of ESBL acquisition versus the acqusition of a particular genotype. B: λ parameter only for genotype models, showing that the estimates are similar for each genotype.

colonisation probability as well as prolonged decay to baseline probability: by the end of the 100 day simulation period in those simulations with both hospital and antibacterial exposure, most probabilities have not yet returned to baseline levels. Shorter course lengths of antibacterials seemed to have similar effects to longer courses. In the model used for the simulations (model two), the effect of all antibacterials (including CPT) is equal and so CPT seems to be the primary driver of an increased long-term carriage probability. TB therapy is also included in the composite "antibacterial" variable, so these conclusions would be equally valid for TB therapy.

8.5 Discussion

In this chapter, I have extended the continuous-time Markov models available in the *msm* package in R to incorporate true time-varying covariates (rather than stepwise constant). I have fitted them to the data presented in Chapter 5 using a Bayesian framework and a differential equation solver in the probabilistic programming language Stan. From these fitted models, it is possible to draw several conclusions.

First, the class of models that I present are feasible to fit in a reasonable amount of time with modest computational requirements, and are very flexible. The models were largely fit overnight on the WSI cluster with four cores and 3GB RAM. These are not particularly onerous requirements, and the times to fit would be expected to be similar on a desktop machine. The parametrisation of the model is extremely flexible; I chose an exponential form of a post-antibacterial effect but any functional form could be used, simply by replacing the function that generates the covariate values, $x(t)$ in Stan model. If a function can be written down, it can be fitted in this framework with minimal effort. This provides, for example, the opportunity to explore *in silico* different hypotheses as to the ways in which antimicrobial exposure drives ESBL-E carriage, by exploring the functional form of the antimicrobial exposure covariate that best fits the data.

Second, the values of the parameter estimates and the simulations from the ESBL models allow an insight into the drivers of ESBL-E colonisation in Malawian adults, and suggests areas to target for interventions. Hospitalisation acts to increase both ESBL-E acquisition and loss, the net result of which is a rapid increase in the probability of ESBL-E colonisation following admission. Antimicrobial exposure acts to reduce the rate of ESBL-E loss and thus prolong carriage, but it appears, from the models, that simultaneous hospitalisation and antimicrobial exposure have a synergistic effect to produce the observed rapid increase in ESBL-E carriage prevalence seen in antibacterial exposed inpatients. This is certainly biologically plausible. The hospital environment at QECH is such that cleaning is difficult,

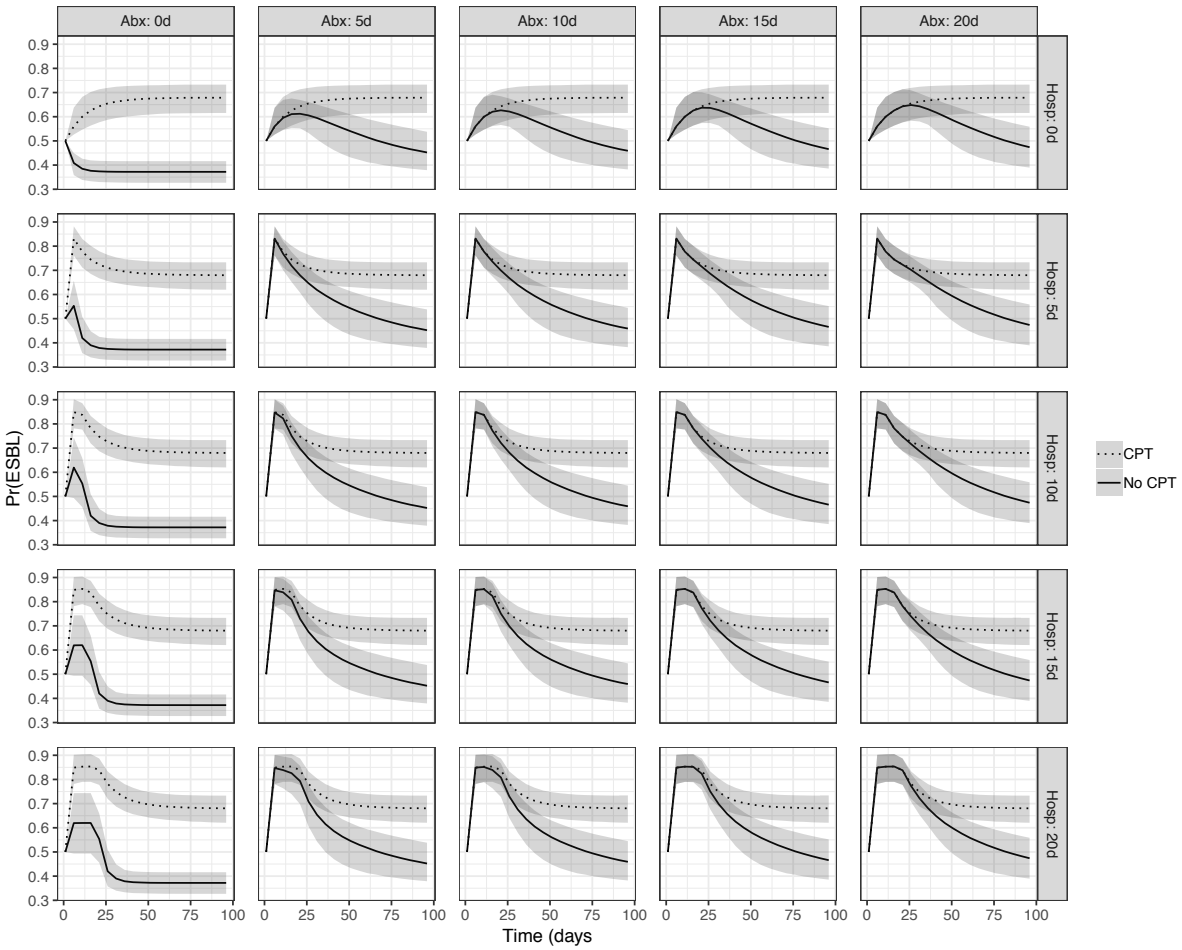

Figure 8.5: Simulations of different antibacterial and hospitalisation scenarios. CPT = Cotrimoxazole preventative therapy. Plots show estimated probability of being in the ESBL+ state for given covariate values as a function of time, assuming a baseline 50% probability of ESBL-E colonisation. Antimicrobial exposure ranges across columns from 1 to 20 days, and hospitalisation across rows from 1 to 20 days. Hospitaliation is clearly the primary driver of rapid initial increase in probability, wheras antimicrobial exposure in the form of CPT is the primary deteriminent of increased long-term carriage probability.

hand washing facilities for staff, participants and guardians are lacking, wards are crowded with participants close together, and one toilet is shared between around 60 patients, all of which potentially facilitate the acquisition of ESBL-E. The genomics data (Chapter 6) suggests that there is no one hospital clone and in terms of *E. coli* diversity at least, the hospital is an extension of the community. Given the number of adult admissions to QECH per year - around 15,000[13], this is perhaps not surprising: if each adult admission is cared for by two guardians then ~5% of the population of Blantyre - 800,000 at the 2018 census - is passing through QECH yearly and the models presented here suggest that QECH may be playing a significant role in driving the high prevalence on ESBL-E carriage in Blantyre. In this situation hospital infection prevention and control (IPC) measures could potentially make a significant impact on the transmission of ESBL-E in Blantyre. Evidence based IPC measures that can be deployed in very resource-limited settings such as QECH are urgently needed.

Strategies to mitigate against the effect of antimicrobials on ESBL-E carriage are also needed. The data presented here support a post-exposure effect of antibacterials on prolonging ESBL-E carriage duration, such that short courses of antimicrobials seem to have a similar effect to longer courses in hospitalised participants. This finding may be contingent on the parametrisation of the post-antibiotic effect, and requires further exploration, but could have significant implications for antimicrobial stewardship. In this model framework, two days of antibacterial therapy to ten inpatients would results in considerably more participant-days carriage of ESBL-E than twenty days of antibacterial therapy to one patient, despite the same number of defined daily doses being used in total. This would suggest that antimicrobial stewardship interventions to avoid unnecessary antibacterials altogether would be more effective than those limiting antibacterial course lengths by e.g. review of blood culture results at 48 hours. The post-antibacterial effect has a lengthy half life of 44 days (95% credible interval 15-98 days), much longer than the time by which most antimicrobials will have been excreted from the body. Such a prolonged effect is biologically plausible, however: even short courses of antimicrobials are known to profoundly alter the composition of the gut microbiota[14,15], which could certainly alter ESBL-E carriage dynamics[16]. Further studies of the role of the microbiota in colonisation resistance to ESBL-E could shed light on the mechanisms of the post-antibacterial effect I demonstrate here, and pave the way for microbiota-modulating therapies to mitigate against it.

The role of CPT in driving long-term ESBL-E carriage is likely significant, and it appears to be a major determinant of long-term ESBL-E colonisation. Again, this is perhaps not surprising given that cotrimoxazole exposure dwarfs exposure to all other antimicrobials combined in the cohort. CPT has been shown to have significant mortality benefits in people living with HIV[17], and lifelong CPT is mandated by Malawian HIV guidelines for all people

living with HIV[18]. Given an estimated adult Malawian HIV prevalence of 9.6%[UNAIDS], CPT is likely therefore a major driver of ESBL-E carriage in Malawi. The risk of driving AMR with CPT needs to be balanced against its benefits, and may be possible that in the era of high ART coverage, reducing malaria incidence and growing Gram-negative resistance that these risks begin to outweigh the benefits. The exact mechanism by which CPT confers a reduced mortality risk - whether it acts primarily to prevent opportunistic infections, bacterial infections or malaria - remains controversial. A recent RCT in Uganda carried out in 2012[19] showed that a strategy of stopping CPT once the CD4 cell count is persistently above 250 cells μL^{-1} is associated with more CPT-preventable infections, including malaria and pneumonia, but no difference in mortality (1.7% vs 1.8% over 12 months). Results are awaited of the TSCQ trial (ClinicalTrials.gov identifier NCT01650558), which has assessed the effect of mortality of CPT versus chloroquine malaria prophylaxis in Malawian HIV-infected adults, based on the hypothesis that in malaria-endemic areas the mortality benefit of CPT is primarily driven by its antimalarial properties. Given the findings here, a chloroquine based prophylaxis strategy could significantly impact ESBL-E carriage prevalence (and hence, possibly, infections) in Malawi and would be very attractive from this perspective if non inferior to CPT in mortality endpoints.

Finally, using WGS as a high resolution typing tool allows very granular insight into ESBL-E carriage at the genotype level. Within the limitations of reasonably uncertain parameter estimates due to small numbers, all parameters for genotype carriage models were the same as the general ESBL carriage model, with the exception of λ. This indicates that the rate of acquisition of a given *E. coli* genotype is two to three orders of magnitude lower then the overall rate of ESBL acquisition, which suggests that apparent continual ESBL-E carriage in fact represents a much more frequent apparent acquisition of different ESBL-E genotypes. This could represent true acquisition or some other dynamic shift in the relative abundance of ESBL producing *E. coli* in the microbiota over time. This analysis is, however, hampered by the fact that only one colony at each time point was sequenced and hence the true distribution of *E. coli* genotypes at a given time point is unknown - see limitations, below.

8.5.1 Limitations

There are several limitations to the analysis presented here. First, despite a reasonable number of data points, the parameter estimates from these models have moderate uncertainty. Some of this may be consequent on the model structure: with strongly correlated parameters, the data may be consistent with a wide range of paired parameter values. Even those parameter values whose 95% credible intervals cross zero (e.g. the hazard ratio of antibacterial exposure on ESBL-E acquisition in model 2) largely incorporate a clinically meaningful effect size, and

so care must be taken not to interpret a lack of certainty of a significant effect as a lack of effect. Uncertainty in parameter estimates increases as more parameters are added, meaning that understanding the relative effects of different antibacterial agents on ESBL-E carriage is not possible, and in most models antibacterials are considered as an aggregate variable. *A priori*, different antibacterial agents would not be expected to have the same effect on ESBL-E carriage dynamics, but here they are considered together. There is some support for this strategy from the fact that the estimated effect sizes of ceftriaxone, ciprofloxacin and amoxicillin (the most commonly administered antibacterials apart from CPT and TB therapy) are similar when considered individually and in aggregate, but uncertainty in these estimates warrants caution. The apparent effect of TB therapy is particularly surprising, given that the first-line combination of rifampicin, pyrizinamide, ethambutol and isoniazid would be expected to have a limited selection pressure for ESBL-E, and warrants further study.

In addition, despite fitting well to participants from arm 1 of the study (those with sepsis), the models fit poorly to arm 2 (antimicrobial unexposed participants) and arm 3 (community members). The reasons for this are not clear, but it strongly suggests that there are covariates that are not included in the model that differentiate the arms of the study in some way. If these covariates are also associated with the exposures of interest (hospitalisation and antibacterial exposure) then this is a potential source of bias from confounding.

These models assume perfect test characteristics, such that the measurement of ESBL-E status (or species or genotype, depending on the model) perfectly represents the true state. This is unlikely to be the case in practice, and there is also likely to be differential test characteristics between the different stool testing methods (stool or rectal swab culture) used. This may have introduced bias to the parameter estimates and simulations. Expanding the model to incorporate imperfect tests - a hidden Markov model - could address this limitation, as well as provide estimates of test sensitivity and specificity. Conceptually this is straightforward; the underlying "true" state is modelled, the likelihood for a given participant ((8.3)) becomes the sum over all possible underlying paths through the system and parameters are added for the sensitivity and specificity of the tests used. This will, however, increase computational costs: if a participant has ESBL-E status measured at n time points then calculating the likelihood required summing over all 2^n possible combinations of states, rather than just one as in the models presented here.

Generalising the model to allow states to be hidden or censored would also address a serious limitation of the genotype models. In these models, the absence of a particular genotype from a sample is interpreted as true absence, but the true situation is more complex. If no ESBL at all is cultured then we can be confident that a given genotype is absent, within the confines of the test sensitivity. However, if *E. coli* were cultured at any time point, then only one

colony pick was taken forward for sequencing, meaning that it is possible that any number of other genotypes were present in the sample but not picked and sequenced and therefore identified. Data on within-participant gut mucosal ESBL-E diversity are sparse, but those data that are available suggest that it may be considerable[20], and so these models should be considered as merely exploratory. Expanding the model to allow states to be censored (i.e. for the true underlying state to remain unknown for a given measurement) is equivalent to the changes that would be necessary to incorporate hidden Markov models, and would address these problems.

8.6 Conclusion and further work

In conclusion, I have developed and fit time-inhomogeneous Markov models to the clinical longitudinal ESBL-E carriage data. The models are computationally tractable, extremely flexible, and provide insight into the drivers of ESBL-E carriage in Blantyre. Though both hospitalisation and antibacterial exposure significantly affect the probability of ESBL-E carriage, they appear to act synergistically together to drive colonisation. Antibacterial exposure seems to have an effect that persists long after most antibacterials would be expected to be excreted from the body; the models provide no data on the mechanism of this but one hypothesis would be that it is mediated by changed in the microbiota. Short courses of antibacterials seem to produce a similar effect to longer courses, which may have implications for antibacterial stewardship interventions. Co-trimoxazole preventative therapy may be one of the major drivers of long-term ESBL-E carriage in Malawian adults and this should be considered in developing international and national guidelines on its use.

These conclusions suggest a direction for future work. The models must be expanded to incorporate censored states to allow the fitting of hidden Markov models and to account for the single colony pick sampling method which was used. This, in conjunction with whole genome sequencing of the remainder of the isolates from the study will allow unbiased models to be fitted to understand carriage at the level of the genotype. Finally, shotgun metagenomic sequencing of stored extracted stool DNA from the participants in this study will a) define the total diversity of ESBL genes within each sample and b) will allow an analysis to identify microbiota associations of ESBL-E colonisation and the effect of antibacterial exposure. This will allow testing of the hypothesis that the post-antibacterial effect I have identified is mediated via the microbiota.

8.7 Appendix

The Stan code for the fitted models is below; the stepwise-constant covariate model is presented fist, and all other models were fitted with the second model. Pairwise posterior parameter estimates for model two (to demonstrate strong parameter correlations) are also shown below; see text for details.

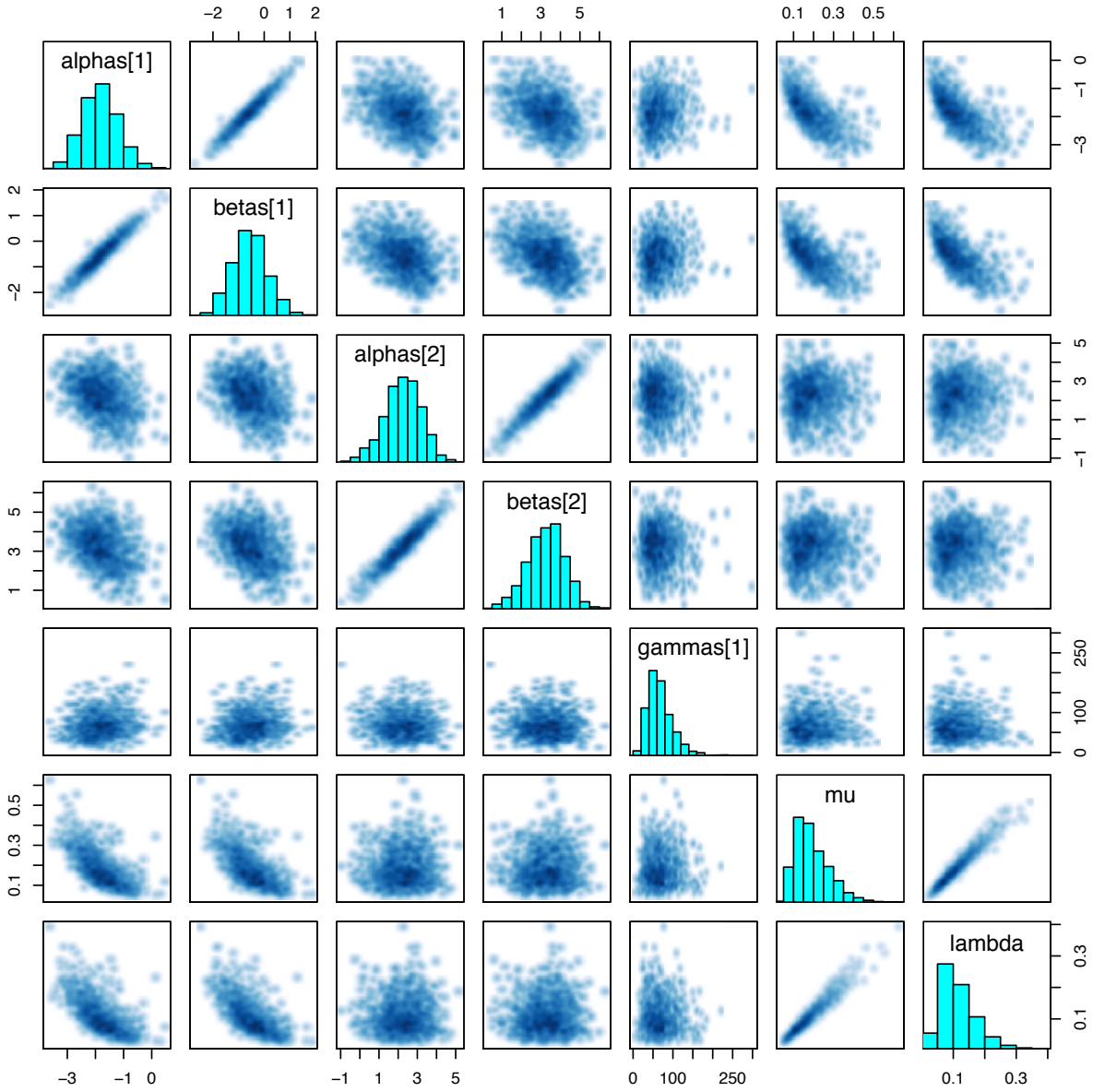

Figure 8.6: ESBL-E carriage model pairwise posterior parameter estimates, showing correlations between alpha and beta for a given covariate, and lambda and mu. These results are for model 2 to predict ESBL-E probability - see text for details. Alpha[1] and beta[1] are the coefficients for the composite antibacterial exposure variable, alpha[2] and beta[2] for hospitalisation, lambda the rate of ESBL-E loss, mu the rate of ESBL-E acquisition, and gamma the scaled (by log(2)) half life of the post-antibacterial effect.

```
// Stan final model incorporating varying numbers of covariates
// Optional gamma decay
// Uses rk45 ODE solver
// Joe Lewis July 2019

// to call this model from Rstan, pass it the following data

// N: integer = number of rows of data, each row consisting of two ESBL
// observations for one patient

// n_covs: integer vector of length 3 = [number of
// nontimevarying covariates,
// number of stepwise constant covariates,
// number of exp decay covariates ]

// covs_type: integer vector of length(number of covariates) =
// each position encodes the type of variable
// in the order they are presented in covs_mat:
    // 3 = time varying with exponential decay of effect
    // 2 = time varying with piecewise constant
    // 1 = nontimevarying
    // All the exp decay variables must always go first

// cov_mat: real matrix of start and stop times of
// covariates 3*(with number of covariate) cols
// Each covariate needs three columns, in this order
    // start_time: time that covariate started
    // stop time: time that covariate stopped.
    // If there is no covariate exposure in this row, code as -999
    // prev_stop time: if covariate has exp. decay, this is
    // the previous stop time (before current row e.g. -10)
    // If no previous exposure, code as 999
    // If non time varying exposure, code as 999 = present, -999 absent

// start state: real vector of length2 = start state in
// format (ESBL-, ESBL+) ie esbl positive coded as
// [0,1] and ESBL negative coded as [1,0]
```

```
// end state: integer length 1, final state.

// this will also generate and save log-likelihoods to do model comparison with loo.

functions {

// Time varying covariate value calculation
// Needs to be passed a 1d array of covariates
// each 3 entries are  (cov_start_time, cov_end_time, prev_cov_end_time)
// prev_cov_end_time is coded as
// t of prev cov end time if has been exposure, pos no if not
// Needs to return a matrix with n_cov rows and 1 column
// to act on the alphas and betas of the model


// n_covs is an array with integer for each cov
// 1 = not tme varying and coded with prev time- present if > 0
// and absent of < 0
// 2 = time varying but no decay; prev time is ignored
// 3 = time varying with decay. If there is no
// exposure in this block, set stop_time to < 0

real[] return_time_varying_coefs_exp_flat(
  real[] cov_mat_passed,
  real t1,
  int[] n_covs_passed,
  real[] gamma_passed
      ) {
        real out_vars[size(n_covs_passed)];
        int s;
        int f;
        int p;

        for (n in 1:size(n_covs_passed)) {
          s = 1 + ((n-1)*3);
          f = s + 1;
          p = f + 1;
               // for each row in cov matrix (ie each covariate)
```

```
if (n_covs_passed[n] == 3) {
     // gamma decay
  if (cov_mat_passed[f] > 0) {   //if there is exposure this block
      if (t1 <= cov_mat_passed[f] && t1 >= cov_mat_passed[s]) {
          // if exposure is happening now
          // set value to 1
      out_vars[n] = 1;
      } else if (t1 > cov_mat_passed[f]) {
          // otherwise if there is exposure in this block
          // and this covariate is set to have a decaying effect
          // and time is after it has stopped
          // set value to decay from stop time
        out_vars[n] = exp((t1-cov_mat_passed[f])/(-1*gamma_passed[n]));
       } else if (t1 < cov_mat_passed[s] && cov_mat_passed[p] < 0) {
         // otherwise, if time is before start time
          // and there is previous exposure
          // set value to decay from previous time
        out_vars[n] = exp((t1-cov_mat_passed[p])/(-1*gamma_passed[n]));
        } else {
           //  otherwise set to 0
      out_vars[n] = 0;
         }
     } else {   // if there is no exposure in this block
         if (cov_mat_passed[p] < 0) {    // if there is previous exposure
           out_vars[n] = exp((t1-cov_mat_passed[p])/(-1*gamma_passed[n]));
         } else {
           out_vars[n] = 0;
         }
      }

    } else if (n_covs_passed[n] == 2) {

      if (t1 <= cov_mat_passed[f] && t1 >= cov_mat_passed[s]) {
          // if exposure is happening now
          // set value to 1
     out_vars[n] = 1;
     } else {
     out_vars[n] = 0;
```

```
            }
           } else if (n_covs_passed[n] == 1) {
             if (cov_mat_passed[p] > 0) {
                out_vars[n] = 1;
             } else {
                out_vars[n] = 0;
             }
           }
      }  // end of for loop
      return out_vars;
      } // end of fn

   // function to return lambda(t) and mu(t)
   // this should take a vector of length n_cov of time
   // varying values of the covariates of the betas
   // (from the time varying coef fn)
   // and two vectors of length n_cov of parameters
   //the alphas (that act on mu)
   // and the betas (that act on lambda)
   // and return a vect or of length two for the
   // values of lamba(t) and mu(t)

 //  real[] return_time_var_transition_hazard(
//      real
 //  )

   // differential state equation

   real[] twostateODE2_flat(real t,   // time
   real[] y,                     // state
   real[] theta,                 // parameters
   real[] x_r,                   //data
   int[] x_i) {                 // data

      // y is state as [p0,p1]
      // theta defined as
      // [ lambda, mu, gamma0, ... gamman,
      //  alpha0, alpha1, ... alphan,
```

```
        //  beta0 ... betan ]
        // where n is number of covariatese
        // data x_r is 1d array of covariates, 3 for each covariate
        // x_i is array of covariate type as
         // [number of non-timedep var,
         //number of timedep nongamma var,
         //number of gamma var,
         // then an integer for each cov:
         //1 (non timedep),2(nongamm) or 3(gamma)]

        real dydt[2];
        real coefs[size(x_i[])-3];  //vector of coefs
        real alphaz[size(x_i[])-3]; // vector of alphas
        real betaz[size(x_i[])-3]; // vector of betas
        real gammaz[x_i[3]];
        real lambda_pr;
        real mu_pr;
        real lambda0;
        real mu0;
        lambda0 = theta[1];
        mu0 = theta[2];
        gammaz =   theta[3:(2+ x_i[3])];
        alphaz = theta[(3+ x_i[3]):(3+x_i[3] + x_i[1] + x_i[2] + x_i[3] -1)] ;
        betaz = theta[(3+x_i[3] + x_i[1] + x_i[2] + x_i[3]):(2+x_i[3] + 2*(x_i[1] + x_i[2]
        coefs = return_time_varying_coefs_exp_flat(x_r, t, x_i[4:size(x_i)], gammaz);
        lambda_pr = lambda0*exp(dot_product(coefs, betaz));
        mu_pr = mu0*exp(dot_product(coefs, alphaz));

        dydt[1] = -y[1]*lambda_pr + y[2]*mu_pr;
        dydt[2] = y[1]*lambda_pr - y[2]*mu_pr;
        return dydt;
    } // end of function
}       // end of block

  data {
      int < lower = 1 > N; // Number of rows of data
      int <lower = 0> n_covs[3];  //[nontimevary, timevarynogamma, timevarygamma]
      int covs_type[sum(n_covs)]; // integer for each cov to define type
```

```
      real t[N];                    // end time
      real cov_mat[N,sum(n_covs[])*3]; // array of covariates, 3 rows for each
      real start_state[N,2]; // start state (at t=0) in form [p0,p1]
      int end_state[N];   // end state (at t) as integer
      }

      transformed data {
         int x_i_pass[3 + sum(n_covs)];
         x_i_pass[] = append_array(n_covs[], covs_type[]);
      }

      parameters {
      real < lower = 0 > lambda;
      real < lower = 0 > mu;
      real <lower = 0> gammas[n_covs[3]];
      real alphas[sum(n_covs[])];
      real betas[sum(n_covs[])];
      }

      transformed parameters {
      real theta[2 + 2*(sum(n_covs)) + n_covs[3]];
      theta[1] = lambda;
      theta[2] = mu;
      theta[3:(2+ n_covs[3])] = gammas[];
      theta[(3+ n_covs[3]):(3+n_covs[3] + sum(n_covs) -1)] = alphas[];
      theta[(3+n_covs[3] + sum(n_covs)):(2+n_covs[3] + 2*(sum(n_covs)))] = betas[];
      }

 model {
      real temp[1,2];
      lambda ~ normal(0,0.2);
      mu ~ normal(0,0.2);
      alphas ~ normal(0,2);
      betas~ normal(0,2);
      gammas ~ normal(0,100);

      for (n in 1:N) {
```

```
        temp = integrate_ode_rk45(twostateODE2_flat,
        start_state[n],
        0, t[n:n],
        theta[],
        cov_mat[n],
        x_i_pass[], 1e-6,1e-6,1e6);

        if (end_state[n] == 1) {
           target += log(temp[1,2]);
        } else {
          target += log(temp[1,1]);
        }
       }
}

generated quantities {
   // needed for loo
    vector[N] log_lik;
    real temp[1,2];
    for (n in 1:N) {
      temp = integrate_ode_rk45(twostateODE2_flat,
      start_state[n],
      0,
      t[n:n],
      theta[],
      cov_mat[n],
      x_i_pass[],
      1e-6,1e-6,1e6);
      if (end_state[n] == 1) {
        log_lik[n] = log(temp[1,2]);
      } else {
       log_lik[n] = log(temp[1,1]);
      }
    }
}
```

```
// Stan model for msm style interval censored model, stepwise constant covariates

      functions {

      // Differential state equations for solving

      real[] twostateODE(real t,        // time
      real[] y,      // state
      real[] theta,  // parameters
      real[] x_r,    // data (real)
      int[] x_i) {   // data (integer)

      real dydt[2];
      real lambda;
      real mu;
      real ab_alpha0;
      real ab_beta0;
      real hosp_alpha1;
      real hosp_beta1;

      real lambda_beta_sum;
      real mu_alpha_sum;

      lambda= theta[1] ;
      mu = theta[2];
      ab_alpha0 = theta[3];
      ab_beta0 = theta[4];
      hosp_alpha1 = theta[5];
      hosp_beta1 = theta[6];

      lambda_beta_sum = 0;
      mu_alpha_sum = 0;

      // first coef, abx, start x_r[1] and end time x_r[2]

      if (x_r[1] == 999) {
      // dont do anything, there is nothing for this covariate
      } else if (t <= x_r[2] && t >= x_r[1]) {
```

```
lambda_beta_sum = lambda_beta_sum + ab_beta0;
mu_alpha_sum = mu_alpha_sum + ab_alpha0;
}

// second coef coef, abx, start x_r[3] and end time x_r[4]

if (x_r[3] == 999) {
// don't do anything, there is nothing for this covariate
} else if (t <= x_r[4] && t >= x_r[3]) {
lambda_beta_sum = lambda_beta_sum + hosp_beta1;
mu_alpha_sum = mu_alpha_sum + hosp_alpha1;
}


dydt[1] = -y[1]*lambda*exp(lambda_beta_sum) + y[2]*mu*exp(mu_alpha_sum);
dydt[2] = y[1]*lambda*exp(lambda_beta_sum) - y[2]*mu*exp(mu_alpha_sum);


return dydt;
}
}

data {
int < lower = 1 > N; // Sample size
real t[N]; // end time
real start_state[N,2]; // start state (at t_start) in form [p0,p1]
int end_state[N];   // end state (at t) as integer
real covariates[N,4]; // covariate start and end times
                    // (as ab_start, ab_end, hosp_start, hosp_end)
}

transformed data {
// real x_r[0];
int x_i[0];

}

parameters {
```

```
real < lower = 0 > lambda;
real < lower = 0 > mu;
real ab_alpha0;
real ab_beta0;
real hosp_alpha1;
real hosp_beta1;
// real < lower = 0 > gamma;
}

transformed parameters {
real theta[6];
theta[1] = lambda;
theta[2] = mu;
theta[3] = ab_alpha0;
theta[4] = ab_beta0;
theta[5] = hosp_alpha1;
theta[6] = hosp_beta1;

}

model {
real temp[1,2];
lambda ~ normal(0,0.2);
mu ~ normal(0,0.2);
ab_alpha0 ~ normal(0,2);
ab_beta0 ~ normal(0,2);
hosp_alpha1 ~ normal(0,2);
hosp_beta1 ~ normal(0,2);
//gamma ~ normal(20,20);
for (n in 1:N) {
temp = integrate_ode_rk45(twostateODE,
start_state[n],
0,
t[n:n],
theta,
covariates[n],
x_i,
1E-6,1E-6, 1E6);
```

```
if (end_state[n] == 1) {
target += log(temp[1,2]);
} else {
target += log(temp[1,1]);
}
}
}

generated quantities {
vector[N] log_lik;
real temp[1,2];
for (n in 1:N) {
temp = integrate_ode_rk45(twostateODE,
start_state[n],
0, t[n:n],
theta,
covariates[n],
x_i,
1E-6,1E-6, 1E6);

if (end_state[n] == 1) {
log_lik[n] = log(temp[1,2]);
} else {
log_lik[n] = log(temp[1,1]);
}
}
}

// The posterior predictive distribution
```

References

1 Teunis PFM, Evers EG, Hengeveld PD *et al.* Time to acquire and lose carriership of ESBL/pAmpC producing E. coli in humans in the Netherlands. *PLOS ONE* 2018;**13**:e0193834. doi:10.1371/journal.pone.0193834

2 Kluytmans-van den Bergh MFQ, Mens SP van, Haverkate MR *et al.* Quantifying Hospital-Acquired Carriage of Extended-Spectrum Beta-Lactamase-Producing Enterobacteriaceae Among Patients in Dutch Hospitals. *Infection Control & Hospital Epidemiology* 2018;**39**:32–9. doi:10.1017/ice.2017.241

3 Haverkate M, Platteel T, Fluit A *et al.* Quantifying within-household transmission of extended-spectrum β-lactamase-producing bacteria. *Clinical Microbiology and Infection* 2017;**23**:46.e1–7. doi:10.1016/j.cmi.2016.08.021

4 Hout A van den. *Multi-state survival models for interval-censored data.* Chapman; Hall/CRC 2016.

5 Longini IM, Clark WS, Byers RH *et al.* Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in medicine* 1989;**8**:831–43.http://www.ncbi.nlm.nih.gov/pubmed/2772443

6 Jackson CH. Multi-State Models for Panel Data: The msm package for R. *Journal of Statistical Software* 2011;**38**:1–28. doi:10.18637/jss.v038.i08

7 Marshall G, Jones RH. Multi-state models and diabetic retinopathy. *Statistics in medicine* 1995;**14**:1975–83.http://www.ncbi.nlm.nih.gov/pubmed/8677398

8 Carpenter B, Gelman A, Hoffman MD *et al.* Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 2017;**76**:1–32. doi:10.18637/jss.v076.i01

9 Gelman A, Carlin JB, Stern HS *et al. Bayesian data analysis.* 3rd ed. Chapman; Hall/CRC 2004.

10 Hoffman MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in

Hamiltonian Monte Carlo. 2014. http://mcmc-jags.sourceforge.net

11 Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 2017;**27**:1413–32. doi:10.1007/s11222-016-9696-4

12 Soetaert K, Petzoldt T, Setzer RW. Solving Differential Equations in R : Package deSolve. *Journal of Statistical Software* 2010;**33**:1–25. doi:10.18637/jss.v033.i09

13 Lewis DK, Callaghan M, Phiri K *et al.* Prevalence and indicators of HIV and AIDS among adults admitted to medical and surgical wards in Blantyre, Malawi. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 2003;**97**:91–6. doi:10.1016/S0035-9203(03)90035-6

14 Palleja A, Mikkelsen KH, Forslund SK *et al.* Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nature Microbiology* 2018;**3**:1255–65. doi:10.1038/s41564-018-0257-9

15 Francino MP. Antibiotics and the Human Gut Microbiome: Dysbioses and Accumulation of Resistances. *Frontiers in Microbiology* 2016;**6**:1543. doi:10.3389/fmicb.2015.01543

16 Buffie CG, Pamer EG. Microbiota-mediated colonization resistance against intestinal pathogens. *Nature reviews Immunology* 2013;**13**:790–801. doi:10.1038/nri3535

17 World Health Organisation. Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach. Second edition. Geneva: 2016.

18 Malawi Ministry of Health. Malawi Guidelines for Clinical Management of HIV in Children and Adults (Third Edition). 2016.

19 Anywaine Z, Levin J, Kasirye R *et al.* Discontinuing cotrimoxazole preventive therapy in HIV-infected adults who are stable on antiretroviral treatment in Uganda (COSTOP): A randomised placebo controlled trial. *PLOS ONE* 2018;**13**:e0206907. doi:10.1371/journal.pone.0206907

20 Stoesser N, Sheppard AE, Moore CE *et al.* Extensive Within-Host Diversity in Fecally Carried Extended-Spectrum-Beta-Lactamase-Producing Escherichia coli Isolates: Implications for Transmission Analyses. *Journal of clinical microbiology* 2015;**53**:2122–31. doi:10.1128/JCM.00378-15