# Developing an Antimicrobial Strategy for Sepsis in Malawi

-

Thesis submitted in accordance with the requirements of the Liverpool School of Tropical Medicine for the degree of Doctor in Philosophy by Joseph Michael Lewis

August 2019

# Contents

**Preface** **9**

**1 Introduction** **11**
- 1.1 Chapter Overview 13
- 1.2 Sepsis in sub-Saharan Africa 13
- 1.3 ESBL-E in sub-Saharan Africa 13
- 1.4 Conclusions 13
- 1.5 Thesis overview 13
- 1.6 Appendix 13
- 1.7 References 13

**2 Methods** **15**
- 2.1 Chapter Overview 17
- 2.2 Study site 17
- 2.3 Clinical Study 17
- 2.4 Diagnostic Laboratory Procedures 17
- 2.5 Molecular methods 17
- 2.6 Bioinformatics 17
- 2.7 Statistical Analysis 17
- 2.8 Study Team 17
- 2.9 Data Collection and Storage 17
- 2.10 Ethical Approval, Consent and Participant Remuneration 17

**3 A clinical and microbiological description of sepsis in Blantyre, Malawi** **19**
- 3.1 Chapter overview 20
- 3.2 Introduction and chapter aims 20
- 3.3 Aims and Methods 20
- 3.4 Results 20
- 3.5 Discussion 20

3.6 Conclusions and further work . . . . . 20
3.7 Appendix . . . . . 20

**4 Exploratory modelling of sepsis outcome** **21**

**5 ESBL-E carriage in Malawian adults in health and disease** **23**
5.1 Chapter Overview . . . . . 24
5.2 Introduction and chapter aims . . . . . 24
5.3 Methods . . . . . 24
5.4 Results . . . . . 24
5.5 Discussion . . . . . 24
5.6 Conclusions and further work . . . . . 24

**6 Whole genome sequencing of ESBL *E. coli* carriage isolates** **25**
6.1 Chapter overview . . . . . 27
6.2 Methods . . . . . 27
6.3 Results . . . . . 27
6.4 Discussion . . . . . 27
6.5 Appendix . . . . . 27

**7 Genomics I** **29**

**8 Continuous time Markov models to understand ESBL-E carriage dynamics** **31**
8.1 Chapter Overview . . . . . 31
8.2 Introduction and chapter aims . . . . . 31
8.3 Methods . . . . . 31
8.4 Results . . . . . 39
8.5 Discussion . . . . . 47
8.6 Conclusion and further work . . . . . 49

**References** **51**

# List of Tables

8.1 Estimates (and standard error) of pairwise expected log pointwise predictive denisty differences for all models . . . . . . . . . . . . . . . . . . . . . . . . . . 40
8.2 Parameter estimates (and 95% confidence intervals) from model 2 . . . . . . . 43

# List of Figures

8.1 Two state ESBL-E model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34
8.2 Parameter estimates from Markov models . . . . . . . . . . . . . . . . . . . . . 41
8.3 Predicted proprtion of ESBL-E positive samples, stratified by arm. . . . . . . 42
8.4 Parameter estimates from models of bacterial species and genotype carriage . 45
8.5 Simulations of different antibacterial and hospitalisation scenarios . . . . . . . 46

# Preface

Placeholder

# Chapter 1

# Introduction

Placeholder

## 1.1 Chapter Overview

## 1.2 Sepsis in sub-Saharan Africa

### 1.2.1 Search strategy

### 1.2.2 Defining sepsis

### 1.2.3 Applicability of sepsis-3 definitions in sub-Saharan Africa

### 1.2.4 Sepsis epidemiology in sub-Sahara Africa

#### 1.2.4.1 Incidence

#### 1.2.4.2 Risk factors: the sepsis population in sub-Saharan Africa

#### 1.2.4.3 Outcomes

### 1.2.5 Sepsis aetiology in sub-Saharan Africa

#### 1.2.5.1 Bacterial zoonoses, Rickettsioses and arboviruses

#### 1.2.5.2 HIV opportunistic infections: PCP, histoplasmosis and cryptococcal disease

### 1.2.6 Sepsis management

#### 1.2.6.1 Early goal directed therapy

#### 1.2.6.2 Evidence to guide antimicrobial therapy in sSA

#### 1.2.6.3 Evidence to guide intravenous fluid therapy in sub-Saharan Africa

## 1.3 ESBL-E in sub-Saharan Africa

### 1.3.1 Search strategy

### 1.3.2 Introduction: definition and classification of ESBL-E

### 1.3.3 Global molecular epidemiology of ESBL-E: an overview

#### 1.3.3.1 1980s-1990s: First identificatiom of ESBL in nosocomial pathogens

#### 1.3.3.2 1990s-2010s: Emergence and globalisation of CTX-M

# Chapter 2

# Methods

Placeholder

## 2.1 Chapter Overview

## 2.2 Study site

### 2.2.1 Malawi

### 2.2.2 Queen Elizabeth Central Hospital

### 2.2.3 Participating Laboratories

#### 2.2.3.1 Malawi-Liverpool-Wellcome Clinical Research Programme

#### 2.2.3.2 Malawi College of Medicine Tuberculosis Laboratory

#### 2.2.3.3 Wellcome Trust Sanger Institute

## 2.3 Clinical Study

### 2.3.1 Entry Criteria

### 2.3.2 Study Visits and Patient Sampling

#### 2.3.2.1 Enrollment assessment and first six hours

#### 2.3.2.2 Subsequent visits

#### 2.3.2.3 Blood, urine, and stool, sputum and CSF collection

#### 2.3.2.4 Imaging: chest x-ray and ultrasound scanning

### 2.3.3 Outcomes and sample size calculations

## 2.4 Diagnostic Laboratory Procedures

### 2.4.1 Point of care diagnostics

### 2.4.2 Laboratory diagnostics

#### 2.4.2.1 Haematology and biochemistry

#### 2.4.2.2 Aerobic blood and CSF culture

#### 2.4.2.3 Mycobacterial blood culture

#### 2.4.2.4 Sputum Xpert

# Chapter 3

# A clinical and microbiological description of sepsis in Blantyre, Malawi

Placeholder

## 3.1 Chapter overview

## 3.2 Introduction and chapter aims

## 3.3 Aims and Methods

## 3.4 Results

### 3.4.1 Study population

### 3.4.2 Symptoms and health-seeking behaviour

### 3.4.3 Admission physiology and laboratory investigations

### 3.4.4 Aetiology

### 3.4.5 Treatment

### 3.4.6 Outcome

### 3.4.7 Determinants of mortality

## 3.5 Discussion

### 3.5.1 Demographics and outcome: significant longer-term mortality

### 3.5.2 Aetiology: TB dominates as a cause of sepsis

### 3.5.3 Determinants of 28-day mortality: an expanded role for TB therapy?

### 3.5.4 Limitations

## 3.6 Conclusions and further work

## 3.7 Appendix

# Chapter 4

# Exploratory modelling of sepsis outcome

# Chapter 5

# ESBL-E carriage in Malawian adults in health and disease

Placeholder

## 5.1 Chapter Overview

## 5.2 Introduction and chapter aims

## 5.3 Methods

## 5.4 Results

### 5.4.1 Study population

### 5.4.2 Exposures during the study period

### 5.4.3 ESBL-E colonisation

### 5.4.4 Associations of ESBL colonisation

## 5.5 Discussion

### 5.5.1 Limitations

## 5.6 Conclusions and further work

# Chapter 6

# Whole genome sequencing of ESBL *E. coli* carriage isolates

Placeholder

## 6.1 Chapter overview

## 6.2 Methods

### 6.2.1 Bioinformatic pipeline

### 6.2.2 Global *E. coli* collection

### 6.2.3 Statistical analysis

## 6.3 Results

### 6.3.1 Samples and quality control

### 6.3.2 Phylogroup, MLST and core genome phylogeny of study isolates

### 6.3.3 Study isolates in a global context

### 6.3.4 Antimicrobial resistance determinants

#### 6.3.4.1 $\beta$-lactam resistance

#### 6.3.4.2 Quinolone resistance

#### 6.3.4.3 Aminoglycoside resistance

#### 6.3.4.4 Chloramphenicol, co-trimoxazole, tetracycline and other resistance determinants

#### 6.3.4.5 Clustering and lineage association of AMR determinants

### 6.3.5 Plasmid replicons

### 6.3.6 Testing metadata associations: SNP distance, hierBAPS sequence clusters and ESBL-clusters

#### 6.3.6.1 Hierarchical BAPS clustering of core gene pseudosequences

#### 6.3.6.2 ESBL-clusters

#### 6.3.6.3 Assessing for healthcare-associated lineages

#### 6.3.6.4 Assessing for within-patient conservation of lineage or MGE

## 6.4 Discussion

# Chapter 7

# Genomics I

# Chapter 8

# Continuous time Markov models to understand ESBL-E carriage dynamics

## 8.1 Chapter Overview

whatevs bru

## 8.2 Introduction and chapter aims

## 8.3 Methods

### 8.3.1 Developing the time-inhomogeneous Markov models

In the broadest sense when constructing a model, our aim is to estimate the most likely values of the parameters of the model, $\theta$, 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 usally 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 intermittententently 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 givenparticipant $n$ 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 patricipant $n$; and so for participant $n$ 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, $\theta$. 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:

$$\prod_{k=2}^{j_n} P(S_n(t^n_k)|S_n(t^n_{k-1}), \theta) \tag{8.2}$$

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

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

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 instantanous 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 (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 $\lambda$ and the hazard of ESBL-E loss to be $\mu$ (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:

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

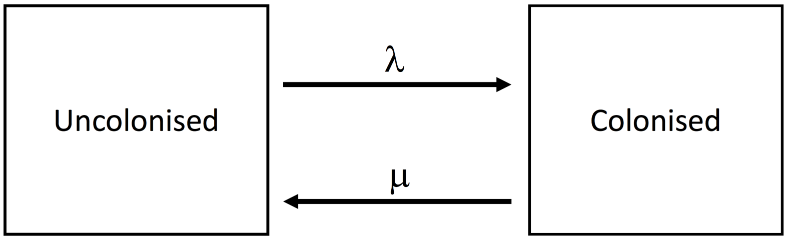

Figure 8.1: Two state ESBL-E model showing instanteneous hazard of ESBL-E acquisition ($\lambda$) or loss ($\mu$).

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 aboev 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 $\theta$. This calculation can be completed for all meaurements 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.

### 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[ref] and the *msm* package in R. 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 $\alpha_k$ and $\beta_k$ are the coefficients of these covariates; these have a straightforward interpretation in that the exponential, $e^{\alpha_k}$ or$e^{\beta_k}$ can be interpreted as a hazard ratio, as per a simple surivival 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 $\gamma_k$ 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 parameterisation has the advantage that the data can fit the size of the parameters $\gamma_k$; if the data are more inkeeping with a stepwise effect of the covariates, then a small ($\ll 1$) $\gamma$ would approximate a step functon 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 $\gamma_k$, and comparing stepwise-constant covariate models to models incorporating the $\gamma_k$ 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 aquisition, respectively; the reciprocals of $\lambda$ and $\mu$ 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 probabablistic programming language *Stan* incorporates an ordinary differential equation solver, and will allow the fitting of the model in a Bayesian framework. 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 distribuion for each parameter. Stan then uses the No-U-Turn Sampler (NUTS) implementation of Markov-chain Monte-Carlo (MCMC) sampling 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 sufficent 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. 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 decsisions 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 variabe 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 $\gamma$ allowed to vary for each agent.

Weakly informative priors were used. A normally distributed prior centered at 0 with standard deviation 2 was used for all the $\alpha$ and $\beta$ 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 centered at 0 with standard deviation 0.2 were used for the $\mu$ and $\lambda$ 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 $\gamma$ 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 macOS 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 disgarding 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 kernal 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. This estimates the out-of sample predictive ability of the model by estimating a quantity called the expected log pointwise predictive denisty ($ELPD$) 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. 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 classifcation obscures a lot of complexity. A participant can be colonised with different ESBL- prodcuing 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 trhough he system. The analysis in Chapter 7 suggests that the element most conserved witin 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 colonistion as a function of time was calculated by solving the equations (8.8) and (8.9) using the R package *deSolve* v1.2.4, 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 cotrimoxazole preventative therapy (CPT).

## 8.4 Results

### 8.4.1 Exploring the effect of antibacterials and hospitalisation on ESBL-E carriage

First, I fit the four models with two aims: to identify the model that provides the best trade off between predictive ability and the computational cost to fit, and to explore the relative effects of hospitalisation versus antimicrobial exposure on ESBL-E carriage by assessing the posterior parameter values of 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. The effect of hospitalisation is consistent across all models; in most models, the 95% credible intervals for both $\alpha_{hospitalisation}$ and $\beta_{hospitalisation}$ 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 addded to the model.

Table 8.1: Estimates (and standard error) of pairwise expected log pointwise predictive denisty 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.

The effect of antibacterial exposure is also reasonably consistent accross the models; the parameter $\alpha$ 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 (incuding 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 $\beta$ 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 acqusition.

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 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 $\alpha$ and $\beta$, the mean time in state for $\lambda$ and $\mu$ and half life of post-antibacterial effect for $\gamma$ are shown in Table 8.2.

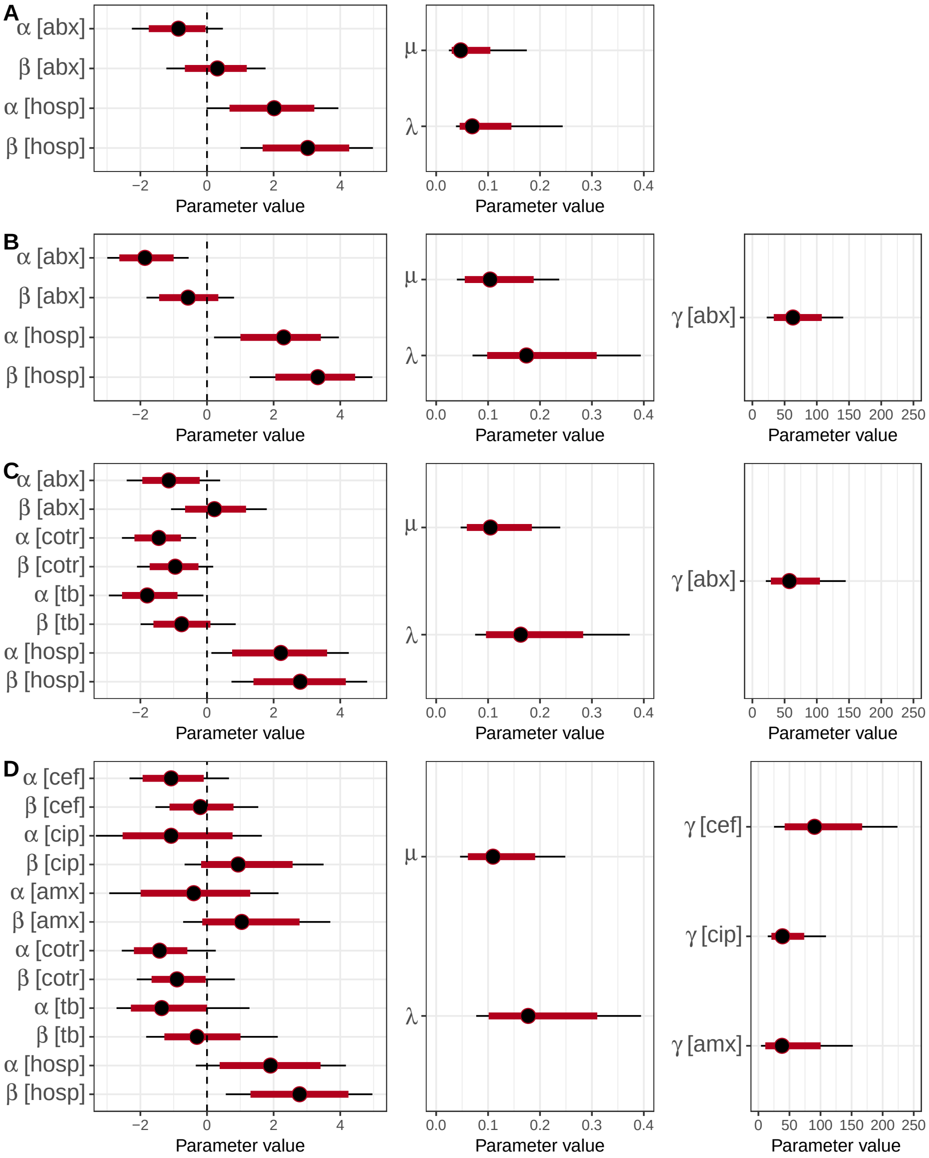

Figure 8.2: Parameter estimates from increasingly complex Markov models to predict ESBL carriage. Black lines are 95% and red lines 80% credible intervals. A: Model 1 includes stepwise constant covariates only, animicrobial exposure and hospitalisation. $\lambda$ is the baseline hazard and $\beta$ the log hazard ratio of ESBL-E acquisition, $\mu$ the baseline hazard and $\alpha$ the log hazard ratio of ESBL-E loss. B: Model 2 adds a post-exposure effect of antimicrobial exposure, parameterised by $\gamma$ 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 $\alpha$[hosp] and $\beta$[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 $\alpha$ 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).

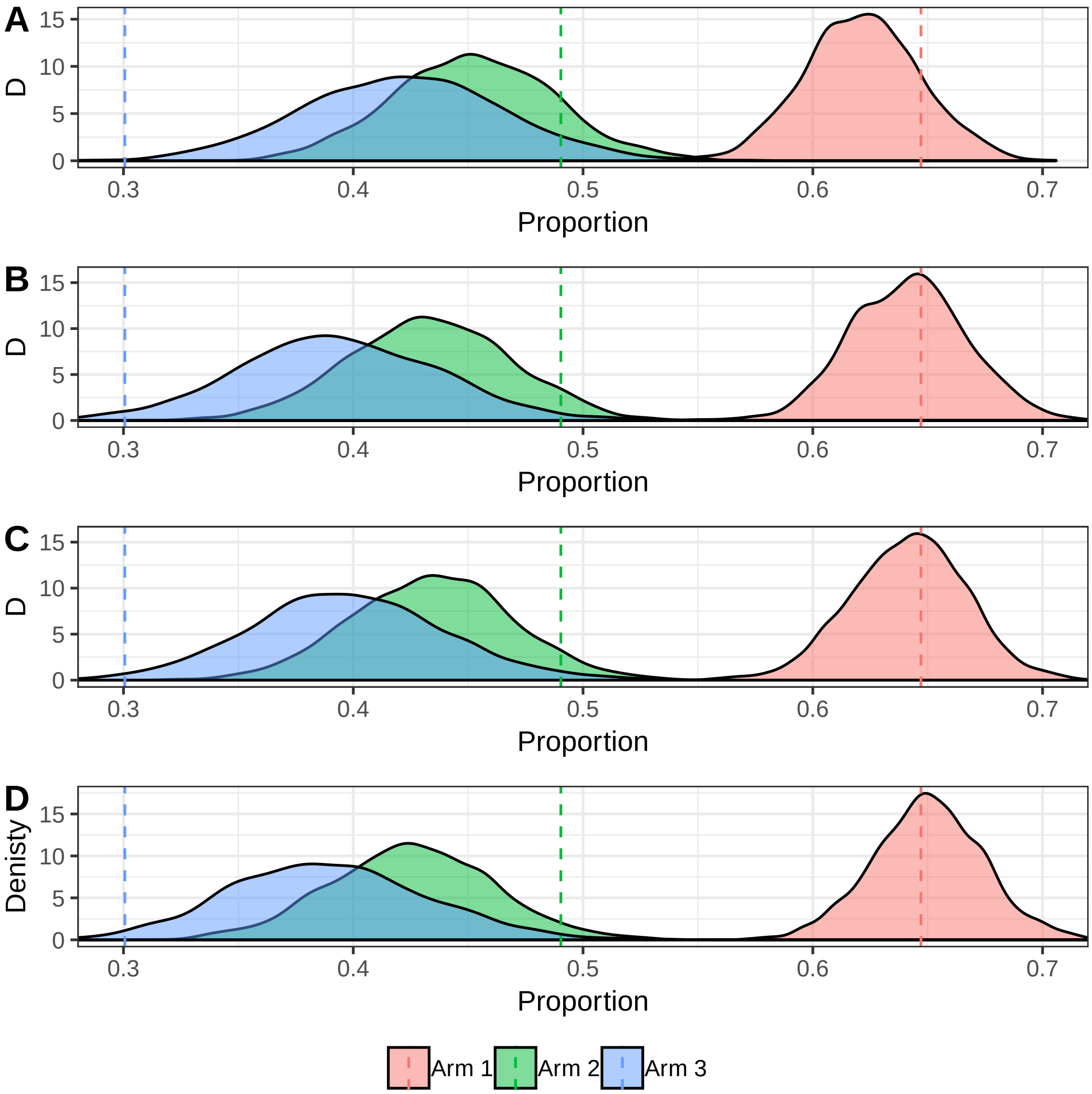

Figure 8.3: Posterior predictive checks: kernal density estimate, D, of predicted proprtion of ESBL-E positive samples, stratified by arm for Model 1 (A), Model 2 (B), Model 3 (C) and Model 4 (D), generated by sampling from a Bernoulli distribution using the predicted probability for each sample (n=993) 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 $\gamma$) 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 $\alpha$ and $\beta$ in the model; half life is equal to log(2) multiplied by $\gamma$; mean time in state assumes all other covariates are equal to zero and is then the reciprocal of $\lambda$ or $\mu$.

### 8.4.2 Exploring bacterial species and gentotype 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 xx]), 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 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 $\lambda$ parameter, the baseline hazard of state acquisition. The magnitude of the difference was large; for example the median (95% CI) $\lambda_{ESBL}$ 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, $\mu$ 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 (CPT). Hospitalisation seems to be the primary determinent of the rapid initial increase in carriage probability, with a lesser effect of antimicrobial exposure, and CPT seems to be the primary driver of an increased long-term carriage probability. In this mode, TB therapy and CPT are included in the composite "antibacterial" variable, so these conclusions would be equally valid for TB therapy.

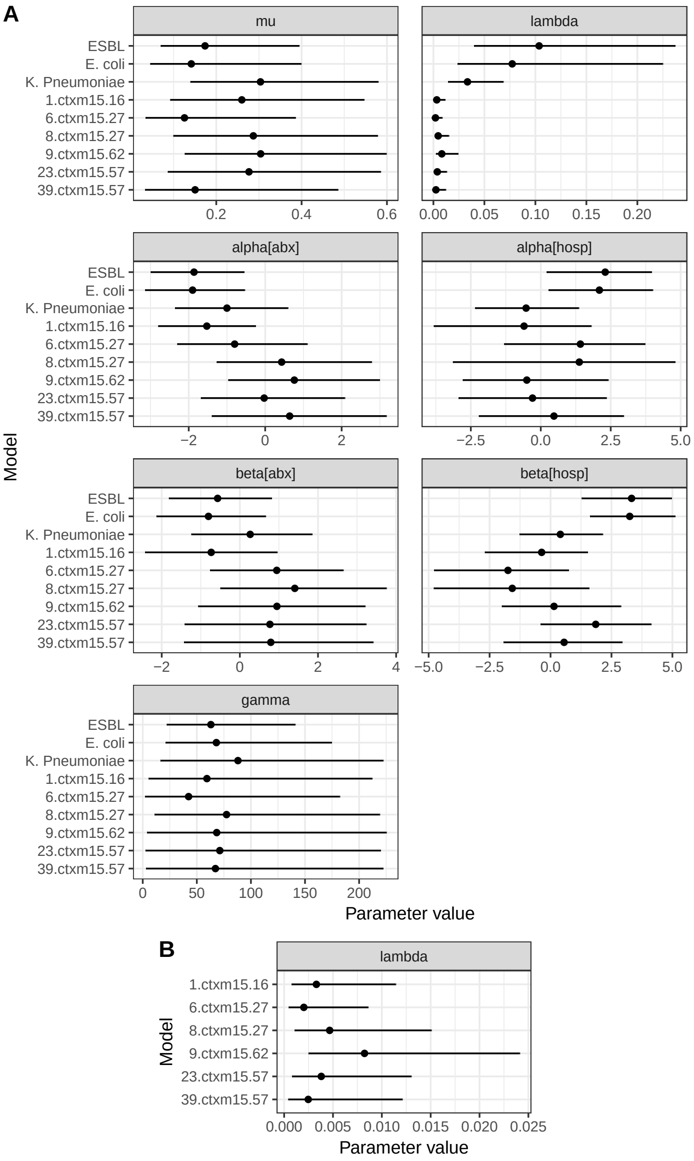

Figure 8.4: Parameter estimates from two state models predicting species and E. coli genotype carriage, compared to original model, which predicted carriahe of any ESBL-E. A: All parameters, showing that the only significant difference between the models is the parameter $\lambda$ (the hazard of acquisition),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

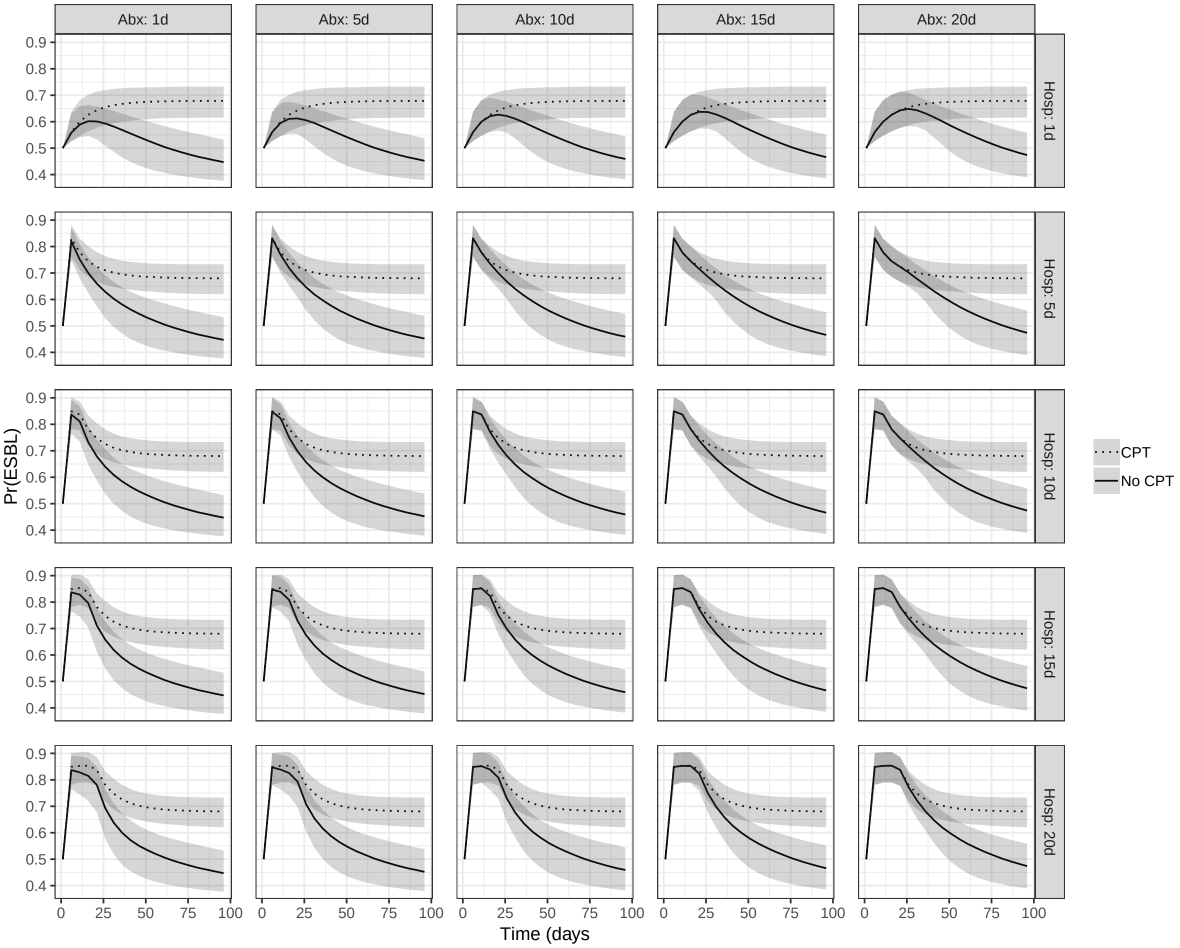

Figure 8.5: Simulations of different antibacterial and hospitalisation scenarios. CPT = Cotrimoxazole preventative therapy. Plots show estimated probability of being in the ESBL+ state for given covariate values as a function of time, assuming a baseline 50% probability of ESBL-E colonisation. Antimicrobial 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.

## 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 ahave fitted them to the data presented in Chapter 5 using a Bayesian framework and a differential equation solver in the probabalistic programming language Stan. From these 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 flexble. 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 parameterisation 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. 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 the dominant driver of rapid increase in ESBL-E prevalence seen in sepsis participants in the study is hospitalisation rather than antibacterial exposure. Some care must be taken with these conclusions both due to the fact that the parameter estimates have wide confidence intervals and the poor fit of arm 2 of the study may indicate unmeasured confounders - see limitations, below - but these associations are certainly biologically plausible. The hospital environment at QECH is such that cleaning is diffcult, hand washing facilites for staff, participants and guardians are lacking, wards are crowded with participants close together, and with one shared toilet 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[ref borgstein HIV paper], 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. Evidnce based IPC measures that can be deployed in very resource-limited settings such as QECH are urgently needed.

The models suggest that antimicrobials administered as treatment for sepsis certainly have an effect on increasing the participant-days of ESBL-E carriage in the cohort, but that this is likely less than the effect of hospitalisation. The role of CPT in driving long-term ESBL-E carriage, on the other hand seems more significant, and appears to be a major determinent 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[ref], and lifelong CPT is mandated by Malawian HIV guidelines for all people living with HIV. Given an estimated adult Malawian HIV prevalence of 9.6%, 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 improves mortality - whether it acts primarily to prevent bacterial infections or malaria - remans controversial. A recent RCT in Uganda in 2012 showed that a strategy of stopping CPT once the CD4 cell count is persistently above 250 cells $\mu L^{-1}$ 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, which has assessed the effect of mortality of CPT versus chloraquine malaria prophylaxis in Malwian HIV-infected adults. Given the findings here, the latter strategy could sigificantly impact ESBL-E carriage prevalence (and hence, possibly, infections) and would be very attractive from this perspective if noninferior to CPT in mortality endpoints.

Third, data support a post-exposure effect of antibacterials on prolonging ESBL-E carriage duration, justifying the choice of time-inhomogenous Markov models to model this system. 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, which could certainly alter ESBL-E carriage dynamics, and immediately sugegst the possibility of microbiota-modulation therapies to mitigate against the effect of antibacterials on ESBL-E carriage. 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 such interventions.

Fourth, 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 $\lambda$. This indicates that the rate of aquisition of a given *E. coli* genotype is two to three orders of magnitude lower then the overall rate of ESBL acquistion, 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

Post antibacterial effect

parameter estimates Some care is warranted in interpreting this finding as a lack of effect of antimicrobial exposure on ESBL-E acquisition, however: firstly, the credible intervals cover a relevent effect size, and secondly, for cotrimoxazole in particular, in many models $\beta$ has a reasonably high probability of being different than zero. For example in Model 3 (Figure 8.1C) the 80% credible interval for $\beta_{cotrimoxazole}$ does not cross zero, meaning there is an 80% probability that $\beta_{cotrimoxazole}$ is different from zero. Nevertheless, given these parameter estimates we can have good confidence in the conclusion that, (conditional on the models and assumptions made) hospitalisation acts to increase both ESBL-E acquisition and loss, and antibacterial exposure acts to reduce ESBL-E loss, and moderate confidence wide confidence intervals incorporate clinically relevent effects even when cross 0

poor fit to arm 2 and 3 - is hospitalisation really the thing or is there confoundig

Wide CIs for cipro, cefo, amoxy - difficult to assess roles of individual agents

models are parameter hungry - more parameters, more uncertainty

Imperfect tests - hidden markov model needed

Censor states - multiple carriage

Interactions - effect of abx different in hospital?

## 8.6 Conclusion and further work

# References