Statistical inference and modelling for nosocomial infections and the incorporation of whole genome sequence data

Colin J. Worby, MMath.

Thesis submitted to The University of Nottingham
for the degree of Doctor of Philosophy

July 2013

Abstract

Healthcare-associated infections (HCAIs) remain a problem worldwide, and can cause severe illness and death. It is estimated that 5-10% of acute-care patients are affected by nosocomial infections in developed countries, with higher levels in developing countries. The increasing level of antibiotic resistance among bacteria that cause HCAIs limits infection treatment options, and is a major concern. This only increases the importance of infection control and prevention methods, particularly in healthcare institutions, where individuals are considerably more susceptible to infection. Hospital infection control policies aim to restrict transmission routes as far as possible.

Statistical modelling is of great importance in increasing understanding of HCAI transmission dynamics. In this thesis, stochastic epidemic models are developed and used with the aim of investigating methicillin-resistant *Staphylococcus aureus* (MRSA) transmission and intervention measures in hospital wards. Bayesian inference allows unobserved transmission dynamics to be taken into account, using a data-augmented Markov chain Monte Carlo algorithm. Using such an approach leads to improved parameter estimation, and allows more flexible and complex models to be analysed. Despite such advantages, much research is still required to develop and assess techniques in this field. Methods to compare models in a Bayesian framework are not well established, and are particularly complex in settings with missing data. Such methods are investigated in this thesis, and a systematic study of Bayesian model choice for transmission models is conducted.

Until recently, large-scale studies which take into account the genetic diversity of the pathogen have not been feasible. Technological advances have meant that the collection of whole genome sequence (WGS) data is now easier, faster and cheaper than ever before, but statistical methods to utilize this information in transmission dynamic models are still in the early stages of development. Analyses relying on routinely collected epidemiological data can derive the rate and relative importance of transmission

routes, but the specific transmission network remains unclear, since all MRSA isolates are necessarily regarded as identical. In this thesis, new methods are developed to model nosocomial MRSA transmission, using genetic information in addition to epidemiological data. Transmission models are constructed which incorporate genetic information where available, where measures of genetic similarity allow us to estimate who infected whom. MRSA transmission was analysed under various assumptions, with the aim of reconstructing transmission networks. Outcomes are compared with a model excluding genetic data, in order to assess the benefits of this new approach.

The collection of WGS data, in combination with new modelling approaches, allows an unprecedented insight into individual level transmission dynamics. This is of much interest to policy makers, as it may aid the investigation of heterogeneity in patient infectiousness and the effectiveness of infection control methods. With WGS data likely to become abundant in the near future, the development of sophisticated analytical tools and models to exploit such genetic information is of great importance.

Acknowledgements

This PhD project was carried out under the supervision of Phil O'Neill, Theo Kypraios, Julie Robotham, Ben Cooper and Daniela De Angelis. I feel lucky to have had such a helpful and encouraging group of supervisors, and am extremely grateful for all of the time, effort and patience that they have dedicated to the completion of this project.

I would like to thank the lovely people in the modelling unit at the Health Protection Agency, with whom I spent the first couple of years of my research. Also, thanks to the other PhD students at the University of Nottingham School of Mathematical Sciences, who were great people to spend a stressful final year with. A special ευχαριστώ to Elena and Sotiris for all the coffee breaks and Greek lessons.

Financial support for this project came from MOSAR, the European network for mastering hospital antimicrobial resistance and its spread into the community.

Finally I'd like to thank my family and friends for being incredibly supportive and understanding, and Emily, who has been patient, reassuring and completely brilliant.

Contents

List of Figures

List of Tables

List of abbreviations

ABC	Approximate Bayesian computation
AIC	Akaike's Information criterion
bp	Base pairs
CI	Confidence interval
CrI	Credible interval
DIC	Deviance information criterion
DNA	Deoxyribonucleic acid
GST	Guy's and St. Thomas' hospital
HCAI	Healthcare-associated infection
HCW	Healthcare worker
ICU	Intensive care unit
MCMC	Markov chain Monte Carlo
MLE	Maximum likelihood estimate
MLST	Multi-locus sequence typing
MPE	Maximum pseudolikelihood estimate
MRSA	Methicillin-resistant *Staphylococcus aureus*
MSSA	Methicillin-susceptible *Staphylococcus aureus*
PCR	Polymerase chain reaction
RJMCMC	Reversible jump Markov chain Monte Carlo
ROC	Receiver operating characteristic
SIR	Susceptible-infectious-removed
SMC	Sequential Monte Carlo
SNP	Single nucleotide polymorphism
VRE	Vancomycin-resistant Enterococci
WGS	Whole genome sequence

CHAPTER 1

Modelling healthcare-associated infections

1.1 Introduction

In this thesis, we explore and evaluate stochastic models to investigate transmission of healthcare-associated infections (HCAIs) in hospitals, and introduce novel methods to integrate whole genome sequence (WGS) data into such models.

We initially describe and work with a framework for the analysis of transmission dynamics using conventional screening data. We utilise Bayesian inference and Markov chain Monte Carlo (MCMC) methods to account for missing data, a common issue in the analysis of epidemics. We investigate the transmission of methicillin-resistant *Staphylococcus aureus* (MRSA), and describe methods to evaluate the effectiveness of infection control measures. It is often of interest to compare transmission models in order to determine the relative support given to a set of candidate models by the data, or to select a best-performing model. The performance of some important tools for Bayesian model choice is investigated in detail. We finally consider the integration of WGS data, with the aim of reconstructing transmission networks. Only recently has the collection of WGS data become feasible on a large scale, and there is much demand for methods to utilise this for the purpose of investigating epidemics and transmission dynamics. Existing methods have limitations and restrictions which make them unsuitable for the analysis of MRSA transmission. Building on the model framework used in earlier work, we develop new methods to estimate the transmission network based on epidemiological and genetic data.

In this chapter, we begin by describing the problem posed by HCAIs, in particular MRSA, in section 1.2. We discuss the important characteristics and transmission mechanisms of this pathogen, and describe infection control measures used by healthcare facilities to reduce the risk of transmission. In section 1.3, a brief overview of the molecular-level dynamics which generate genetic diversity is provided, and methods of collecting genetic data for *S. aureus* are described. In section 1.4, we discuss the fundamental concepts of deterministic and stochastic epidemic models, with an emphasis on methods which account for partially-observed data. In section 1.5, we discuss methods of parameter inference, initially describing likelihood optimisation methods, before discussing Bayesian inference. We describe MCMC methods which will be used throughout the thesis. An overview of Bayesian model selection methods is provided in section 1.6. In section 1.7, we describe the existing literature for modelling HCAI transmission, and review existing methods to integrate epidemiological and genetic data into the analysis of pathogen transmission. We provide some concluding remarks in section 1.8, before describing the aims and structure of the following chapters in section 1.9.

1.2 Healthcare-associated infections

1.2.1 Background

Healthcare-associated infections are a major cause of increased morbidity and mortality in healthcare facilities, often requiring costly treatment and extended hospital stays [1]. Hospital-associated bloodstream infections are the 10th most common cause of death in the USA [2], while it has been estimated that there are in excess of four million cases of HCAI each year in Europe, to which 37,000 deaths can be attributed [3]. HCAIs are estimated to affect 5-10% of hospital admissions in industrialised countries, but are likely to be a larger problem in resource-limited nations, where infection rates have been estimated to range between 6-27% [4]. Infections amongst newborns are 3-20 times more common than in industrialised countries, and are associated with a mortality rate of 50% [5].

Antimicrobial resistance in HCAIs has been increasing over the last decades. Antibiotic effectiveness has been likened to a finite natural resource, which is 'used up' through the administration of antibiotic treatment [6]. As pathogens develop increasing drug-resistance, treatment options become limited. The potential loss of effectiveness of an-

tibiotics has enormous consequences, and it is therefore of great interest to fully understand the dynamics of pathogen transmission, in order to focus on the prevention, as well as the treatment, of infections.

1.2.2 *Staphylococcus aureus*

Staphylococcus aureus is a gram-positive bacterium, asymptomatically carried by around 20% of the population at most times (persistent carriers), and by around 60% intermittently [7], with some geographic fluctuation [8]. The most common carriage site is the anterior nares (nose), although *S. aureus* is also commonly found on the skin (particularly broken skin/wounds), groin, perineum, urinary tract and pharynx [8–11]. *S. aureus* can also enter the bloodstream (bacteraemia), which can cause a number of complications, and may put the individual at risk of severe illness or death. Around 20% of bloodstream infections in US hospitals are caused by *S. aureus* [2].

1.2.2.1 Antibiotic resistance

After its discovery in 1928, penicillin was used to treat *S. aureus* infections, but resistance emerged rapidly in the next decades. The first penicillinase-producing *S. aureus* isolates (resistant to the effects of penicillin) were described in 1944, and within a few years, most hospital isolates were penicillin-resistant [12].

Methicillin was introduced in 1959, and used to treat penicillin-resistant *S. aureus* infections, but resistance developed rapidly, as with penicillin. Just two years later, the first cases of methicillin resistance were reported in the UK [13]. The prevalence of methicillin resistance increased in subsequent decades; in 2001, 47.3% of *S. aureus* samples were resistant in the UK [14], and over 50% in the USA in 2003 [15]. Similarly, strains resistant to other antibiotics and antiseptics commonly used to treat infections have emerged [16–18]. With the emergence of resistance to vancomycin, [19, 20] an antibiotic often considered to be the drug of 'last resort' [13], the threat of pathogens resistant to all available antibiotics has become clear. The careful administration of antibiotics is necessary to slow the proliferation of multiply-resistant pathogens such as *S. aureus* [21].

Methicillin-resistant *Staphylococcus aureus* (MRSA) encompasses *S. aureus* strains resistant to methicillin, as well as practically all beta-lactam antibiotics [22]. Studies into the differences in the epidemiology of MRSA and methicillin-susceptible *S. au-*

reus (MSSA) revealed that the mortality rate in persons infected by the resistant type is higher [23, 24]. The length of stay associated with a patient acquiring MRSA bacteraemia is significantly longer than that associated with MSSA bacteraemia, and has a higher economical impact [6, 25]. However, this is not necessarily a causal effect — compared to those with MSSA infections, patients with an MRSA infection are more likely to have been hospitalised for longer, previously had surgery, or suffered from other medical issues [26].

1.2.2.2 Transmission, carriage and infection

It is widely acknowledged that most *S. aureus* transmission between patients occurs indirectly, via the hands of healthcare workers [27–29]. A systematic review of healthcare worker colonisation concluded that healthcare workers (HCWs) played an important role in the transmission of MRSA, as a vector, rather than a source [29]. Of the studies included in this review which used genotyping to match patient carriage with HCW carriage, 93% found evidence for the existence this transmission route.

Individuals vary in susceptibility to *S. aureus* carriage and infection. Risk increases with age [30], which is also the most consistent predictor for *S. aureus* infection mortality [31, 32]. In addition, previous hospitalisation, skin lesions and use of indwelling devices are risk factors associated with MRSA carriage [31, 33, 34].

A distinction can made between 'colonisation' and 'infection' for *S. aureus* carriage; the former may indicate carriage without clinical symptoms of infection [35]. Throughout this thesis, we define colonised patients to be carriers of MRSA, regardless of infection symptoms.

S. aureus infections can vary in severity, from relatively minor skin and soft tissue infections, to potentially fatal bacteraemia [36], although carriage is most commonly asymptomatic [12]. Infections may be the result of acquisition from another individual, or via organisms which have been present on the host for some time. It has been shown that asymptomatic nasal carriage of *S. aureus* can be a source of bacteraemia in hospitalised patients [37]; indeed, it was found that 80% of patients with bacteraemia are infected with a strain matching the type previously carried asymptomatically [38].

1.2.2.3 MRSA outside of healthcare facilities

While MRSA is most commonly associated with hospitals (denoted hospital-acquired MRSA (HA-MRSA) when a distinction should be made), there are several other sources of colonization or infection which are of importance. MRSA outbreaks have been observed amongst prisoners, athletes, homeless persons and intravenous drug users [39]. Outbreaks of MRSA occurring outside of a healthcare setting are termed community-acquired MRSA (CA-MRSA). CA-MRSA has been found to cause infections in individuals lacking the risk factors associated with HA-MRSA infection [40–42], and is frequently carried at different body sites [42]. While community strains are typically susceptible to a wider range of antibiotics [40, 43], it is feared that strains are becoming more resistant [44]. CA-MRSA is a growing problem, with few established infection control guidelines [45].

1.2.3 Infection control measures

1.2.3.1 Hand hygiene

As HCW contact is widely believed to be the primary route of transmission, elimination of transient carriage from HCWs between patient contacts is key to reducing transmission via this route. Simulation models have confirmed this effect [27], and increased hand hygiene compliance has been shown to be associated with decreased MRSA infection rates [46]. A systematic review revealed that the rate of compliance (that is, hand washing before and after contact with patients) is typically low (overall median of 40%), and that it was lower in an ICU setting than elsewhere in a hospital [47].

1.2.3.2 Screening on admission

While not a control measure itself, screening patients for presence of communicable pathogens such as MRSA on admission to a hospital ward reduces the risk of asymptomatic carriage going undetected. A patient is usually swabbed at one or more of the typical carriage sites (nose, axilla, perineum, groin) for the presence of MRSA. Screening enables the prompt implementation of reactive infection control measures for MRSA positive patients, and in combination with an effective intervention, may in theory contribute to the reduction in transmission. Mandatory screening was advised by the Department of Health in England in 2009 for all elective patients on admission

to hospital, extending to all emergency admissions by the end of 2010 [48]. However, studies on the impact of screening have shown mixed results [49]. Screening has traditionally been carried out with the use of culture swabs, although polymerase chain reaction (PCR) tests are becoming cheaper, and can provide results much more quickly. PCR tests provide results within 24 hours, compared to the two to three days necessary for culture methods, but are considerably more expensive [50, 51]. There has been much debate surrounding the use of rapid detection methods as standard to screen for MRSA carriage. While the introduction of PCR screening can reduce the number of undetected positive patient days, a systematic review in 2009 concluded that the reduction in MRSA acquisition associated with the use of rapid screening was not significant [51].

1.2.3.3 Patient isolation

Patient isolation of some form is a core component of most infection control policies [52]. It has been recommended that patients are isolated in a single side room if resources permit, or in a cohort with other positive patients [21, 53]. The implementation of isolation precautions for MRSA carriers has been considered controversial due to the lack of robust supporting evidence of its effectiveness [54], and the cost implementations associated with its use [55]. There has been little formal evidence to support the use of barrier precautions and physical isolation in reducing MRSA transmission rates [54, 56].

The systematic review of available evidence for isolation usage conducted by Cooper et al. [56] in 2003 concluded that there was a lack of well designed studies in this area. The authors developed a dynamic transmission model to theoretically evaluate isolation, and demonstrated the importance of isolation capacity and timing to the success or failure of isolation in reducing transmission. Since then, Cepeda et al. [57] undertook a prospective trial during which MRSA positive patients were not physically isolated for a period of six months in two intensive care units in London. This period was then compared to a control phase in which isolation in a side room or cohort was used. Standard precautions were maintained throughout. Acquisition rates were found to be similar in both periods, and there was no evidence to suggest increased transmission during the period with no isolation.

Forrester et al. published two studies on a dataset collected from an ICU in Brisbane, using different model-based analyses, which both indicated a beneficial effect of isola-

tion, albeit with a considerable degree of uncertainty [58, 59]. More recently, Kypraios et al. [52] conducted another model-based evaluation of barrier precautions, using data collected in several ICUs in Boston. They found some evidence to support the use of isolation; a best estimate of 28% reduction in transmission was given, with a high degree of uncertainty.

1.2.3.4 Contact pattern limitations

Since MRSA transmission is driven by patient-HCW interaction, a potential intervention would be to limit the contact of susceptible patients with HCWs who have attended colonised patients. Ueno and Masuda presented a model-based analysis of contact limitation to reduce HCAI transmission [60]. The authors used an epidemic model based on a contact network to determine the role of patient-HCW interaction in transmission, and investigated different contact patters. They demonstrated that a reduction may be achieved by assigning patients to a particular team of HCWs. Milazzo et al. performed a simulation-based study to determine the roles of spatial and staff cohorting to reduce transmission of MRSA [61]. They concluded that changing staff contact patterns had a large impact in reducing transmission events — physical isolation/separation of patients alone may not be sufficient. While such contact pattern restrictions may be theoretically advantageous, such an implementation in reality may be infeasible and costly. Staff deficiencies, which result in a higher workload and contact rate for HCWs, have also been identified as being associated with increased MRSA transmission [62].

1.3 Whole genome sequence data

Until very recently, the collection of whole genome sequence (WGS) data has been prohibitively complex and expensive. However, technological advances and falling costs mean that DNA sequencing is now feasible on a larger scale, and it is likely that such data will become abundant in the future. WGS data is of great interest in the study of population dynamics and evolution [63, 64]. For disease-causing organisms, it can provide an insight into the molecular-level mechanisms behind the development of pathogenicity and drug resistance [38, 64]. Both within-host and between-host genetic dynamics may be investigated [63–65]. From an epidemiological perspective, genetic data can aid the analysis of the geographic propagation of a pathogen, as well pro-

viding insight into competition between different strains [66]. Older molecular typing methods may differentiate between different strains of an organism, but lack the resolution to analyse the genetic diversity occurring on an individual level; so-called microevolution [66]. The ability to observe and quantify such diversity allows us to estimate person-to-person transmission routes with unprecedented accuracy.

1.3.1 The analysis of genetic data

All organisms possess a genetic signature encoded in the genome. The genome of a bacterium is typically a single, circular chromosome; a double-stranded DNA structure [67]. DNA comprises a chain of components, known as base pairs (bp), each of which is a pair of nucleotides; cytosine (C) and guanine (G), or adenine (A) and thymine (T). Since each nucleotide may bond with only one other type to form a base pair (C with G, and A with T), information is duplicated in the two strands, and it suffices to consider only one when comparing two isolates. A full DNA sequence, or WGS, may be represented as a vector X of length L, where each element $X_i \in \{C, A, G, T\}$ represents a nucleotide.

Genome length varies by species. Bacterial genomes range from 580-9105 kilobases (kb, equal to 1000 bp) [67]. In comparison, viral genomes are typically much smaller [38]; for instance, human immunodeficiency virus (HIV) has a genome length of 2.5-9kb, influenza A is 13.9kb [68]. Simpler genomes are easier and faster to sequence, and as such, much of the earlier work using sequencing data examined viruses (eg. [69–71]).

Bacteria reproduce via binary fission, a process in which a cell's DNA is replicated and the cell divides. The replication of DNA is not always perfect. Single point mutations cause a base pair to be incorrectly replicated in the 'daughter' cell. Point mutation can be modelled as a Markov chain, with a rate matrix Q, where $Q_{i,j}$ gives the rate at which a nucleotide i mutates to nucleotide j. The probabilities of particular changes may vary, for instance, it is considered that transitions ($A \leftrightarrow G$ or $C \leftrightarrow T$) are more likely than transversions (any other nucleotide change). Various mutation models exist, such as the Jukes-Cantor model, in which all mutation rates are identical, the Kimura model, where transitions and transversions take different rates, and the generalised model in which all substitution types may occur at different rates [72]. The overall rate of point mutations may vary across different species.

Genetic changes on a larger scale may also occur. Insertion and deletion of genetic material (usually a small number of base pairs) can occur during replication, and result in

a larger or smaller genome respectively. Horizontal gene transfer is the acquisition of genetic material from sources other than the parent (vertical transfer). In bacteria, this can occur via processes called transformation, transduction and conjugation. Transformation is the uptake and incorporation of DNA from the environment ('free DNA'). Transduction is the acquisition of genetic material from a virus (bacteriophage). Conjugation occurs during the interaction between two cells, in which plasmids may be exchanged [73]. A plasmid is a DNA molecule existing within a cell, but is not associated with the chromosome, and replicates independently [74].

Horizontal gene transfer can cause substantial changes in the DNA of an organism, and can result in changes in characteristics and behaviour. It is considered likely that this is of great importance to the development of antibiotic resistance in bacteria [75, 76]. Antibiotic resistance may be acquired via the transfer of such genes on plasmids, and the speed of adoption of such complex traits has been found to be inconsistent with the accumulation of single point mutations [77].

Genetic distance between isolates may be measured by the number of point differences, or single nucleotide polymorphisms (SNPs). This can be used as a measure of relatedness, or similarity. Furthermore, this may indicate how recently two organisms diverged. The molecular clock hypothesis states that SNPs accumulate at an approximately constant rate over time [78], allowing time since divergence to be estimated.

1.3.2 Molecular typing of *Staphylococcus aureus*

Methods to categorise and differentiate between strains of *S. aureus* based on molecular typing methods have emerged and developed in the last two decades. Whole genome sequencing is the identification of the complete genome, which in the case of *S. aureus*, is 2.8-2.9 million base pairs (Mb) in length [79]. This, until recently, has been infeasible due to high costs and the intensive processing required. Many existing studies on the transmission of infectious diseases have used viral pathogens, which have a much shorter genome length than bacteria.

Earlier methods in genetic typing of bacteria have been based on technologies which target specific genes or regions of the DNA. Pulsed field gel electrophoresis (PFGE) was first introduced as a method to discriminate between MRSA strains in 1991 [10, 80]. PFGE is a process by which a chromosome is broken down into large segments by an enzyme. Since genetically identical strains are always broken down in the same way, the lengths of the resulting DNA fragments can be used to discriminate between differ-

ent types. PFGE was long considered to be the gold standard of bacterial typing [81]. *spa* typing is a method developed in the 1990s, in which a region of the protein A gene is sequenced [82, 83], and has been used to identify over 7000 *S. aureus* types in total [84]. Multilocus sequence typing (MLST) involves the partial sequencing of the full DNA sequence. DNA fragments from a small set of housekeeping genes are sequenced to create an allelic profile for each *S. aureus* isolate [85]. While these methods allow us to broadly categorise strain types, the resolution is not sufficient to investigate small-scale genetic differences, or 'microevolution', which is required to study transmission over a short period of time [66, 81].

1.4 Epidemic modelling

Statistical models may be used to estimate parameter values and uncertainty based on observed data, as well as to predict the behaviour of a system under certain scenarios. Modelling has been used for decades to gain insight to the transmission process of communicable diseases.

1.4.1 Compartmental models

Epidemics are commonly described using compartmental models. In this approach, we define a number of disease states and model how the number of individuals in each state changes over time. Perhaps the most well-known epidemic model is the susceptible-infectious-removed (SIR) model. The SIR model is characterised by three disease states; susceptible (S), infected (I), and removed (R), the latter of which may be interpreted as death, immunity, or physical isolation. An effective contact between an infectious and a susceptible person results in a transmission event. The definition of an effective contact can depend on the typical routes of transmission and contagiousness of a particular pathogen. Infectious individuals are removed from the population through death or recovery and the development of immunity. These individuals play no further role in the dynamics of the epidemic. A diagram of such a model is shown in figure 1.1.

Pathogens have a diverse range of characteristics, meaning that many alternative models of this type have been created which may be more appropriate. Such models include:

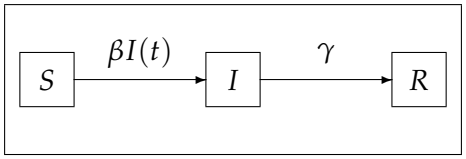

Figure 1.1: The individual-level SIR model for a closed population. A susceptible individual becomes infected at a rate $\beta I(t)$ at time t, where $I(t)$ is the number of infective individuals at time t. Infected individuals are removed, through death or recovery, at rate γ.

- SI — infected individuals are not removed, and remain infectious.
- SIS — infected individuals recover, and are immediately susceptible again .
- SIRS — infected individuals recover, and are immune to infection for a period, before becoming susceptible again due to waning/loss of immunity.
- SEIR — there exists a latent, or 'exposed', period between time of infection and time of infectiousness (during which individuals are in compartment 'E'). Infectious individuals are eventually removed, and are not susceptible to infection again.

1.4.2 Transmission

Transmission of a pathogen may occur via different mechanisms. For instance, proximity to an infected individual may be sufficient for transmission of an airborne pathogen, whereas others require direct contact. Vector-borne diseases are those which typically require the intermediate carriage of a third party to transmit the pathogen (for example, mosquitoes in the case of malaria, and transiently-colonised healthcare workers in the case of some nosocomial pathogens such as MRSA).

The rate of transmission also depends on the contact pattern of individuals within the population, the susceptibility of individuals at risk of infection, and the infectiousness of carriers. In the simplest case, we may assume that individuals mix homogeneously (so that any pairs of individuals have the same chance of interaction), and that susceptibility and infectiousness are the same for all susceptible and infectious individuals respectively. The risk of infection in a fixed-size population is then proportional to the number of infectious individuals at any given time, so that the rate of transmission at time t is $q(t) = \beta S(t)I(t)$, where β is the transmission coefficient, incorporating the contact rate and probability of transmission per contact, and $S(t)$ and $I(t)$ are the

number of susceptible and infectious individuals at time t respectively. This follows the law of mass action [86]. In a population of variable size (including, for example, immigration and emigration), transmission may be dependent on the number, or proportion, of infectious individuals in the population, representing density-dependent ($q(t) = \beta S(t)I(t)$) or frequency-dependent mass action ($q(t) = \beta S(t)I(t)/N(t)$) respectively. There has been some debate as to which formulation is more appropriate in certain settings [87, 88]. Frequency dependent transmission is generally more appropriate for scenarios in which the contact rate does not increase with the population size [89].

In reality, the assumption of homogeneous mixing may not be appropriate. Contact patterns and contact types vary considerably between different age groups and social settings, which can impact model outcomes considerably if not taken into account [90]. Heterogeneous mixing approaches have been developed, accommodating local and global level transmission [91]. Such approaches demonstrate the importance of accounting for heterogeneous mixing when appropriate.

Epidemic models can be described by either deterministic or stochastic dynamics. While this thesis is primarily concerned with the construction and analysis of stochastic epidemic models, we briefly introduce both here.

1.4.3 Deterministic models

In a deterministic epidemic model, the number of individuals in each compartment at any time is specified by the initial state of the model, and the model parameters which describe transmission dynamics. The standard closed population deterministic SIR model is governed by the system of differential equations,

$$\begin{aligned} \frac{\mathrm{d}S(t)}{\mathrm{d}t} &= -\beta S(t)I(t) \\ \frac{\mathrm{d}I(t)}{\mathrm{d}t} &= \beta S(t)I(t) - \gamma I(t) \\ \frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \gamma I(t), \end{aligned} \tag{1.4.1}$$

where β is the effective contact rate between individuals, and γ is the rate of removal of infected individuals. Initial conditions $S(0) > 0, I(0) > 0, R(0) = 0$ must also be defined. This is a special case of the model described by Kermack and McKendrick [92], who are widely considered to have laid the foundation for modern epidemic modelling [93]. This deterministic model describes disease prevalence in a closed population of

N individuals ($S(t) + I(t) + R(t) = N$), where individuals mix randomly and homogeneously, and there is no variation in susceptibility, or the expected duration of infection. This model is easily adapted and extended to model more complex dynamics.

Models defined by a system of differential equations, such as the SIR model, describe integer-valued population counts as continuously varying values. This has little impact for large populations, however, this is inappropriate for small counts.

1.4.4 Stochastic models

Unlike the fixed values realised in a deterministic model, a stochastic model describes a probability distribution for outcomes. In a compartmental model, the number of individuals in each compartment at a given time is modelled as a probability distribution. Such a model allows random fluctuations to play a role in the structure of an epidemic — under identical parameters and conditions, an epidemic could grow large, or die out quickly.

The stochastic version of the SIR model is governed by the equations

$$\begin{aligned} P[S(t+\delta t) = s-1,\, I(t+\delta t) = i+1 | S(t) = s,\, I(t) = i] &= \beta si\delta t + o(\delta t) \\ P[S(t+\delta t) = s,\, I(t+\delta t) = i-1 | S(t) = s,\, I(t) = i] &= \gamma i\delta t + o(\delta t) \\ P[S(t+\delta t) = s,\, I(t+\delta t) = i | S(t) = s,\, I(t) = i] &= 1 - \beta si\delta t - \gamma i\delta t + o(\delta t), \end{aligned} \tag{1.4.2}$$

which describe the probability of moving from the current state of the model, $(S(t), I(t))$, to the state $(S(t+\delta t), I(t+\delta t))$ in the next short period of time δt, a period short enough such that the probability of multiple transition events is small [94, 95].

1.4.4.1 Likelihood for the SIR model

Parameter estimation is of central importance throughout this thesis, so we now consider the likelihood function for a realisation of the SIR model. We suppose the SIR epidemic is perfectly observed, so that the sets of n ordered infection times $\tau^I = \{\tau_1^I, \ldots, \tau_n^I\}$, and m removal times $\tau^R = \{\tau_1^R, \ldots, \tau_m^R\}$ are known with certainty. We suppose that individuals mix homogeneously, and we define the rate of infection, as $q(t) = \beta S(t) I(t)$. For an epidemic to occur, we must begin with at least one infected individual in a population containing susceptible individuals. We assume here that there is initially one infective person, with an infection time at $\tau_1^I = t_0$.

We partition the time interval $[t_0, T]$ by the infection and removal event times $t_0 = e_1 < \cdots < e_{m+n} \leq T$, such that on each interval $[e_j, e_{j+1})$, the rate of infection is constant. The probability of infection occurring at time t depends only on the time of the previous event t', and the number of susceptible and infected individuals at that time $S(t')$, $I(t')$.

We use results from survival analysis to construct the likelihood function. We define the hazard rate, $h(t)$, as the rate at which events (either infection events or removal events) occur. This is the sum of the infection rate and the rate of removal at time t, that is

$$h(t) = \beta S(t) I(t) + \gamma I(t). \tag{1.4.3}$$

The survival function, $A(t_0, t)$, is the probability that no such events occur in the interval (t_0, t). It has been shown (eg. [96]) that this may be expressed in terms of the hazard rate as

$$A(t_0, t) = \exp(-\int_{t_0}^{t} h(u)\mathrm{d}u). \tag{1.4.4}$$

The hazard rate is piecewise constant on the intervals between each of the events $[e_1, e_2), \ldots, [e_{m+n-1}, e_{m+n})$, which means that for all j,

$$\int_{e_j}^{e_{j+1}} h(u)\mathrm{d}u = h(e_j)(e_{j+1} - e_j). \tag{1.4.5}$$

The likelihood of observing the set of infection times $\tau_2^I, \ldots, \tau_n^I$ and removal times $\tau_1^R, \ldots, \tau_m^R$ can then be considered to be the product of contributions representing each infection or removal event, and the preceding interval in which no event occurred. Each event is dependent only on the previous event time. Given the initial conditions $S(t_0)$ and $I(t_0) = 1$, the likelihood is then

$$\begin{aligned}
\pi(\tau^I, \tau^R | \beta, \gamma, S(t_0), \tau_1^I) &= \underbrace{\prod_{i=1}^{m+n-1} A(e_i, e_{i+1})}_{\text{no events occur}} \underbrace{\prod_{j=2}^{n} q(\tau_{j-}^I)}_{\text{infection events}} \underbrace{\prod_{k=1}^{m} \gamma I(\tau_{k-}^R)}_{\text{removal events}} \\
&= \prod_{i=1}^{m+n-1} \exp\big(-\int_{e_i}^{e_{i+1}} h(u)\mathrm{d}u\big) \prod_{j=2}^{n} q(\tau_{j-}^I) \prod_{k=1}^{m} \gamma I(\tau_{k-}^R) \\
&= \exp\big(-\sum_{i=1}^{m+n-1} (\beta S(e_i) I(e_i) + \gamma I(e_i))(e_{i+1} - e_i)\big) \\
&\quad \cdot \prod_{j=2}^{n} \beta S(\tau_{j-}^I) I(\tau_{j-}^I) \prod_{k=1}^{m} \gamma I(\tau_{k-}^R) \\
&= \exp\big(-\int_{t_0}^{T} \beta S(t) I(t) + \gamma I(t)\mathrm{d}t\big) \prod_{j=2}^{n} \beta S(\tau_{j-}^I) I(\tau_{j-}^I) \prod_{k=1}^{m} \gamma I(\tau_{k-}^R),
\end{aligned} \tag{1.4.6}$$

where $\tau^I_{j-} = \lim_{t \uparrow t^I_j} t$, the time just prior to τ^I_j [94, 95].

In the above case, infection and removal times are observed, but not linked to particular individuals. Under the Markov assumption that the probability of each event depends only on the last, it suffices to know only the infection times and removal times, without this link. However, if infection times and removal times are known for each individual, the likelihood may be rewritten as the product of contributions for each individual. This can be useful when considering heterogeneity in terms of infectiousness or susceptibility. We denote the infection time for individual j as t^I_j. If the individual does not become infected, we set $t^I_j = \infty$. Similarly, t^R_j is the removal time of individual j, and is set to $t^R_j = \infty$ if no removal event takes place. We define t^I and t^R to be the vectors of infection times and removal times respectively. In this case, we consider the rate of infection for a given susceptible individual, which is $q(t) = \beta I(t)$. Suppose we have a closed population of N individuals. Let the first infected person be labelled 1, $t^I_1 = t_0$. The likelihood of observing infection and removal times t^I and t^R in a closed population, given the first infection time, is then

$$\begin{aligned}
\pi(t^I, t^R | \beta, \gamma, t_0) = & \prod_{i=1}^{m+n-1} \Big[\prod_{j: t^I_j > e_i} \exp(-\beta I(e_i)(e_{i+1} - e_i)) \prod_{k: t^I_k \le e_i < t^R_k} \exp(-\gamma(e_{i+1} - e_i)) \Big] \\
& \cdot \prod_{j=2}^{n} \beta I(t^I_{j-}) \prod_{k=1}^{m} \gamma \\
= & \prod_{j=1}^{N} \Big[\prod_{i: e_i < t^I_j} \exp(-\beta I(e_i)(e_{i+1} - e_i)) \\
& \cdot \Big(\mathbf{1}_{t^I_j = \infty} + \mathbf{1}_{t^I_j \neq \infty} \prod_{k: t^I_j < e_k \le t^R_j} \exp\{-\gamma(e_{k+1} - e_k)\} \beta I(t^I_{j-}) \Big) \gamma^{\mathbf{1}_{t^R_j \neq \infty}} \Big] \\
= & \prod_{j=1}^{N} \Big[\exp\Big(- \int_{t_0}^{\min(t^I_j, T)} \beta I(u) \mathrm{d}u \Big) \\
& \cdot \Big(\mathbf{1}_{t^I_j = \infty} + \mathbf{1}_{t^I_j \neq \infty} \exp\{-\gamma(\min(T, t^R_j) - t^I_j)\} \beta I(t^I_{j-}) \Big) \gamma^{\mathbf{1}_{t^R_j \neq \infty}} \Big], \quad (1.4.7)
\end{aligned}$$

where $\mathbf{1}_x$ is the indicator function, equal to 1 if x holds, and zero otherwise.

1.4.4.2 Discrete-time

The likelihood for a perfectly observed SIR epidemic is easily adapted for the discrete-time case. Infections occurring in a given time interval (which we define as a day here, for convenience) are assumed to be independent, and the newly-infected individual does not contribute to the number of infectious individuals until the start of the subse-

quent interval, that is,

$$I(t+1) = \text{No. infected on day } t \text{ or earlier} - \text{No. removed on day } t \text{ or earlier.}$$

Furthermore, susceptibles are independent in terms of the avoidance of infection. The probability of a given susceptible individual avoiding infection on day d is $\exp(-\beta I(d))$, where $I(d)$ is the number of infective individuals on day d. Newly infected individuals contribute to the infective population from the day after infection. Similarly, a given susceptible acquires infection on day d with probability $1 - \exp(-\beta I(d))$.

The discrete case can be derived from the continuous version by supposing that events occur at the start/end of each day, so that the hazard rate is piecewise-constant for each day.

Using discrete time is a reasonable approximation when the rate of infection is low and the time intervals are short. Evaluation of the likelihood is also less computationally expensive than in the continuous-time case.

1.4.4.3 Chain binomial models

In the discrete-time framework above, the number of infected individuals on day $t+1$ is

$$I(t+1) = I(t) + X_t - Y_t,$$

where X_t is the number of new infections on day t, and Y_t is the number of infectious individuals removed on day t. Further, given the number of susceptible individuals on day t, X_t follows a binomial distribution

$$X_t | S(t) \sim \text{Bin}(S(t), 1 - \exp(-\beta I(t))).$$

If the infectious period is of a fixed length, then the infectious dynamics are completely specified by this binomial distribution.

Such a model is an example of a chain binomial epidemic model. The Reed-Frost model is a simple, well-known chain binomial model, in which the number of susceptible individuals at time $t+1$ is dependent only on the number of susceptible and infected individuals at time t [86]. In this model, the latent period and duration of infectiousness are assumed to last one unit of time, or one 'generation', and the probability of effective contact between an infectious and susceptible individual is p. As such, each susceptible avoids infection with probability $(1-p)^{I(t)}$, and the number of susceptibles at time

$t+1$ is modelled as a binomially distributed random variable

$$S(t+1) \sim \text{Bin}(S(t), (1-p)^{I(t)}),$$

while the number of infectious individuals is simply $I(t+1) = S(t) - S(t+1)$.

1.4.5 Basic reproduction number, R_0

The transmission potential of a pathogen can be represented by the basic reproduction number of the pathogen, R_0. This is the expected number of secondary infections from a typical infected individual in a large, completely susceptible, population [89, 97, 98]. This dimensionless value depends on both the pathogen (infectiousness, infectious period, etc.) and population (susceptibility, contact pattern, etc.) under consideration. R_0 is a key quantity in identifying epidemic control measures and optimal vaccination strategies.

For a deterministic SIR model, epidemic dynamics can be determined by the value of R_0. An epidemic will occur, that is, the number of infectives increases, if $R_0 > 1$, otherwise the epidemic will die out. In a stochastic implementation of the SIR model, with an infinite initial susceptible population, there is a positive probability that infinitely many susceptibles become infected if $R_0 > 1$, otherwise, the outbreak is certain to be finite. In a finite population, roughly speaking, an outbreak is expected to be small if $R_0 \leq 1$, while a large outbreak can occur if $R_0 > 1$. If the transmissibility of a pathogen remains constant during the infectious period (of mean length L) and the chance of infectious contact is homogeneous throughout the population (at rate β), then $R_0 = \beta L$.

In a hospital outbreak setting, this measure is difficult to evaluate, since the population is small, not entirely susceptible, and, in the case of MRSA, individuals are typically discharged before loss of carriage. Cooper et al. described the net single admission reproduction number as a more suitable measure of the transmissibility of a nosocomial pathogen, and defined it as the average number of secondary cases generated during a single ward admission (episode), where not everyone is necessarily susceptible [99].

1.5 Parameter inference

There are several statistical techniques to derive parameter estimates from observed data under a given model. In this thesis, we primarily employ Bayesian inference techniques. However, we now provide an overview of some of the important tools for

parameter inference from both classical and Bayesian statistics.

1.5.1 Likelihood-based inference

The likelihood of observed data x, given a model specified by parameters $\theta \in \Theta$, is denoted $\pi(x|\theta)$. The maximum likelihood estimate $\hat{\theta}$ is derived through maximisation of the likelihood function (or, equivalently, the log-likelihood function). The maximum likelihood estimate (MLE) of a likelihood $\pi(x|\theta)$ is then

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \pi(x|\theta).$$

1.5.1.1 Optimisation

While in simple cases, maximum likelihood estimates may be derived analytically, this is not generally possible — instead, iterative optimisation procedures must be used to approximate maximal points.

The Newton-Raphson method sequentially approaches optimal points where the derivative is equal to zero, and requires the first and second derivatives (or numerical approximations). Alternatively, the Nelder-Mead simplex algorithm is a derivative-free method, in which a simplex is repeatedly moved around the parameter space, using expansion, contraction and reflection moves to approach a maximum point [100]. This method is advantageous when derivatives are complex or not calculable, but, as with the Newton-Raphson approach, may get stuck in local maxima. Repeating the algorithm with several different starting points can reduce the risk of this. An alternative method is simulated annealing, which incorporates moves to 'worse' (lower likelihood) points while exploring the parameter space, allowing the possibility to leave local maxima [101].

The likelihood function for the fully-observed SIR model described earlier (equation (1.4.6)) may be maximised in order to derive estimates for β (transmission rate) and γ (removal rate);

$$\hat{\beta} = \frac{n-1}{\int_{t_0}^{T} S(u)I(u)\mathrm{d}u},$$
$$\hat{\gamma} = \frac{m}{\int_{t_0}^{T} I(u)\mathrm{d}u},$$

where n and m are the total number of infections and recoveries respectively. However, if transmission and recovery times are not perfectly observed, the likelihood becomes

intractable, and an alternative approach must be used to derive estimates.

1.5.1.2 EM algorithm

The expectation-maximisation (EM) algorithm allows parameters to be estimated in a situation where the likelihood, $\pi(x|\theta)$, is intractable due to unobserved data. The method, introduced by Dempster et al. [102], may be applied in cases where the incorporation of a set of latent data z, in addition to the observed data x, makes the full likelihood $\pi(x, z|\theta)$ tractable. The algorithm is described below:

Expectation-Maximisation algorithm

1. Set initial parameter values, $\theta^{(0)}$.
2. Calculate the expectation of the log likelihood of the full data, $\log \pi(x, z|\theta)$, with respect to z, conditional on the data x and current parameter estimates $\theta^{(i)}$,
$$E(\log \pi(x, z|\theta)|x, \theta^{(i)}). \tag{1.5.1}$$
3. Maximise this expectation with respect to θ, to obtain $\theta^{(i+1)}$
4. Repeat steps 2 and 3 until subsequent iterations are below a given similarity threshold.

While the likelihood $\pi(x|\theta)$ cannot be assessed directly, it can be shown that repeatedly maximising the expectation shown in (1.5.1) with respect to θ generates a chain of estimates which have increasing likelihoods [102, 103]. The algorithm may be alternatively reformulated as two repeated maximisation steps, firstly over the parameters θ, and then over the latent data z [104].

The EM algorithm has been criticised for slow convergence [103], but accelerated versions have been described (eg. [105]). In addition, convergence to saddle points or local maxima is also possible. As with other optimisation techniques, repeating the process with different starting points can reduce the risk of this.

This approach can be useful in the study of epidemics, as data are typically only partially observed. Becker provided an application of the EM algorithm, using a simple

example of a household disease outbreak, described by a chain binomial model [106]. This approach becomes complex for large households, and for more sophisticated models.

1.5.1.3 Uncertainty

Likelihood optimisation returns a point estimate for parameters, but does not indicate a level of uncertainty. Confidence intervals can be calculated in various ways.

Consider a model with parameters θ, with a log-likelihood function of $\ell(\theta)$. Denote θ_{-i} to be the set of parameters excluding θ_i. The profile log-likelihood of a parameter θ_i is given as

$$\ell(\theta_i) = \max_{\theta_{-i}}(\ell(\theta_i, \theta_{-i})),$$

that is, the maximised likelihood where θ_i is fixed [107]. The profile likelihood describes a submodel nested within the full model. The log ratio of the maximised likelihood to the profile likelihood at a point θ^* asymptotically follows a χ^2 distribution under the null hypothesis that $\theta_i = \theta^*$. This fact may be used to construct profile likelihood confidence intervals [108].

The Hessian matrix $H(f)$ for a function $f(\theta_1, \ldots, \theta_m)$ is defined as the $m \times m$ matrix of second order partial derivatives of f, such that

$$H(f)_{i,j} = \frac{\partial^2 f(\theta_1, \ldots, \theta_m)}{\partial\theta_i \partial\theta_j}.$$

The Hessian matrix evaluated at x describes the multidimensional curvature of the function at that point. Intuitively, low curvature of a likelihood surface at the MLE corresponds to a large degree of uncertainty surrounding the estimate.

Asymptotic normality of the MLE $\hat{\theta}$ means the covariance matrix is given by the inverse of the negative Hessian matrix, evaluated at the MLE [109].

Parametric bootstrap methods provide a simulation-based approach to calculating confidence intervals, as well as assessing functions of parameters [110]. Suppose data $x = \{x_1, \ldots, x_n\}$ are a set of realisations from a distribution function $F(x, \theta)$. If $\hat{\theta}$ is an estimate of model parameters θ, then we can draw parametric bootstrap samples $\tilde{x}^{(1)}, \ldots, \tilde{x}^{(m)}$ from the distribution $F(x, \hat{\theta})$. In doing so, we can obtain a set of estimates $\hat{\theta}^{(1)}, \ldots \hat{\theta}^{(m)}$, which may be used to approximate the distribution of θ for sufficiently large m [111]. A version of the bootstrap procedure runs as follows:

Parametric bootstrap

1. Derive a maximum likelihood estimate $\hat{\theta}$ using observed data x.
2. Simulate a set of observations $\tilde{x}^{(i)}$ under the model, given $\hat{\theta}$.
3. Derive a maximum likelihood estimate $\tilde{\theta}^{(i)}$, using the simulated data $\tilde{x}^{(i)}$ and $\hat{\theta}$.
4. Repeat steps 2 and 3 to obtain m samples from the bootstrap distribution.

Having drawn samples from the bootstrap distribution, there are a number of ways to derive a confidence interval. The percentile method is a simple approach to generating a $100(1-\alpha)\%$ confidence interval; this involves taking the $100\alpha/2$ and $100(1-\alpha/2)$ percentiles of the sample [112]. This can result in a biased confidence interval if the distribution is not symmetric about the true value θ [113]. Efron described a bias correcting version to estimate confidence intervals [114]. A review of methods to create confidence intervals is provided by DiCiccio and Efron [115].

1.5.2 Bayesian inference

1.5.2.1 Introduction

Bayesian inference provides a framework in which parameters are viewed as having probability distributions rather than fixed values, as is the case in frequentist analyses [116]. Such inference is based around Bayes' theorem, which, for data x and parameters $\theta \in \Theta$, is

$$\begin{aligned}\pi(\theta|x) &= \frac{\pi(x|\theta)\pi(\theta)}{\pi(x)} = \frac{\pi(x|\theta)\pi(\theta)}{\int_\Theta \pi(x|\theta)\pi(\theta)\mathrm{d}\theta} \\ &\propto \pi(x|\theta)\pi(\theta), \end{aligned} \tag{1.5.2}$$

where $\pi(\theta|x)$ is the posterior density of θ, $\pi(x|\theta)$ is the likelihood of x, and $\pi(\theta)$ is the prior density of θ.

Bayesian inference relies on the specification of prior distributions for parameters θ. These should be selected carefully, as poorly chosen prior distributions can greatly af-

fect the posterior estimates, as can be seen in Bayes' theorem (1.5.2). Prior distributions may be chosen to reflect the existing knowledge and uncertainty about the parameters in question, and may be derived using previous studies or expert knowledge. Alternatively, uninformative prior distributions are chosen when one requires estimates driven almost entirely by the data rather than prior beliefs, either due to lack of existing evidence for parameter values, or to a desire to derive estimates based solely by the observed data.

We aim to evaluate the posterior distribution of parameters θ. Features of this distribution (mean, median, variance etc.) can be expressed as the posterior expectation of functions of θ;

$$E(f(\theta)|x) = \frac{\int_\Theta f(\theta)\pi(x|\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta \pi(x|\theta)\pi(\theta)\mathrm{d}\theta}. \tag{1.5.3}$$

However, the integration required to evaluate such distributions is often intractable [117]. Markov chain Monte Carlo (MCMC) methods to generate samples from posterior distributions had their origins in the 1950s, but until the development of computational technology towards the end of last century, such methods remained impractical for most problems [117].

1.5.2.2 Markov chain Monte Carlo

Monte Carlo integration offers a method by which expectations may be approximated by repeated sampling. For random samples $X_1, \ldots, X_n$ from some distribution $\pi(\cdot)$, the expectation $E_\pi(f(X))$ may be approximated as

$$E_\pi(f(X)) \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i). \tag{1.5.4}$$

However, if π is non-standard, as is typically the case with posterior densities, it is not possible to draw independent samples. Instead, dependent samples may be drawn from a Markov chain, with stationary distribution $\pi(\cdot)$ [118]. A Markov chain is a sequence of random samples $X_1, \ldots, X_n$, where each sample X_j is generated from a transition kernel, $P(X_j|X_{j-1})$, dependent only on the previous sample X_{j-1}. In order to evaluate a posterior distribution $\pi(\theta|x)$, we desire a chain such that, after a number of iterations k, subsequent points represent a sample from this distribution. Various algorithms exist to generate such chains.

1.5.2.3 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a method to generate a Markov chain $\theta^{(1)}, \ldots, \theta^{(N)}$, which, for large enough N, converges to a target distribution $\pi(\cdot)$. At each iteration, a candidate point, θ^* is sampled from a proposal density $q(\theta^*|\theta^{(i)})$, which gives the probability density of proposing θ^*, given the current, ith value. The algorithm runs as follows:

Metropolis-Hastings algorithm

1. Set initial parameter values, $\theta^{(0)}$, and number of iterations, N.

2. Sample a new parameter value, θ^* randomly from the proposal probability density $q(\theta^*|\theta^{(i)})$.

3. With probability

$$\alpha(\theta^*, \theta^{(i)}) = \min\left(1, \frac{q(\theta^{(i)}|\theta^*)\pi(\theta^*)}{q(\theta^*|\theta^{(i)})\pi(\theta^{(i)})}\right), \tag{1.5.5}$$

accept the proposed point, and set $\theta^{(i+1)} = \theta^*$, else set $\theta^{(i+1)} = \theta^{(i)}$.

4. If $i < N$, go to step 2.

From equation (1.5.5), it can be seen that

$$\frac{\alpha(\theta^*, \theta^{(i)})}{\alpha(\theta^{(i)}, \theta^*)} = \frac{q(\theta^{(i)}|\theta^*)\pi(\theta^*)}{q(\theta^*|\theta^{(i)})\pi(\theta^{(i)})}$$

$$q(\theta^*|\theta^{(i)})\alpha(\theta^*, \theta^{(i)})\pi(\theta^{(i)}) = q(\theta^{(i)}|\theta^*)\alpha(\theta^{(i)}, \theta^*)\pi(\theta^*),$$

which satisfies the detailed balance equation, so the stationary distribution of the Markov chain, if it converges, is the target distribution $\pi(\theta)$ [118]. Full justification that the Metropolis-Hastings algorithm produces a chain which converges to the target distribution can be found, for example, in [119].

The Metropolis-Hastings algorithm will always eventually converge to the target distribution, provided all points in the support of this distribution may be proposed. The choice of the proposal function is nevertheless of great importance to the speed of convergence, and efficiency of the algorithm, which we look at in closer detail in section 1.5.2.4.

The Metropolis algorithm is a special case of the Metropolis-Hastings algorithm, in which the proposal distribution is symmetrical; that is, $q(X|Y) = q(Y|X)$. This simplifies the acceptance probability (1.5.5) to

$$\alpha(\theta^*, \theta^{(i)}) = \min\left(1, \frac{\pi(\theta^*)}{\pi(\theta^{(i)})}\right).$$

The Metropolis-Hastings algorithm is commonly used to estimate the posterior distribution of parameters θ, given observed data x. In this case, the target distribution is $\pi(\theta|x) \propto \pi(x|\theta)\pi(\theta)$. The normalising constant in the posterior distribution cancels in the acceptance ratio, meaning that it suffices to evaluate the likelihood and prior functions for proposed points θ^*.

1.5.2.4 Convergence & mixing

The choice of proposal density $q(\cdot)$ in the Metropolis-Hastings algorithm is of great importance, and affects the rate at which the samples converge to the target distribution, as well as the rate at which the distribution is explored (mixing).

If the proposal samples are very close to the current value, then the acceptance rate is likely to be high, but the chain will take longer to converge to the target distribution, and will only mix slowly (that is, the sampled points will only gradually move around the parameter space). Conversely, if the proposal steps are too large, then a move is unlikely, and the acceptance rate will be low. Figure 1.2 shows the effect of the proposal distribution on the mixing and convergence of the algorithm, using a simple example. We attempt to sample from a normal target distribution of $N(1, 1)$, using a normal proposal distribution centred on the current value, $\theta^{(i)}$. The proposal variance σ^2 is assigned values of 0.01, 1, and 100. The lowest variance exhibits a high acceptance rate (98%), but very slow convergence — we can see that the two independent chains do not converge after 10000 iterations. With a variance of $\sigma^2 = 1$, acceptance is 60%, and convergence is quick. Running the algorithm with $\sigma^2 = 100$ results in very slow mixing and low acceptance (1%). While the Metropolis algorithm will converge for any proposal distribution, it is clear that a careful choice of σ^2 is important for efficient sampling. The acceptance rate can be 'tuned' to give a desired rate of acceptance by altering the variance of the proposal distribution. This may be specified before running the algorithm, possibly via investigation of pilot runs to determine optimal acceptance rates. Roberts et al. proposed an optimal acceptance rate of 0.234 for Metropolis algorithms, under fairly general conditions [120]. The proposal distribution may also be

adapted during the algorithm. For example, with a normal proposal distribution, the acceptance rate over the last k iterations may be measured, and the variance adapted as necessary, in order to increase or decrease the acceptance rate.

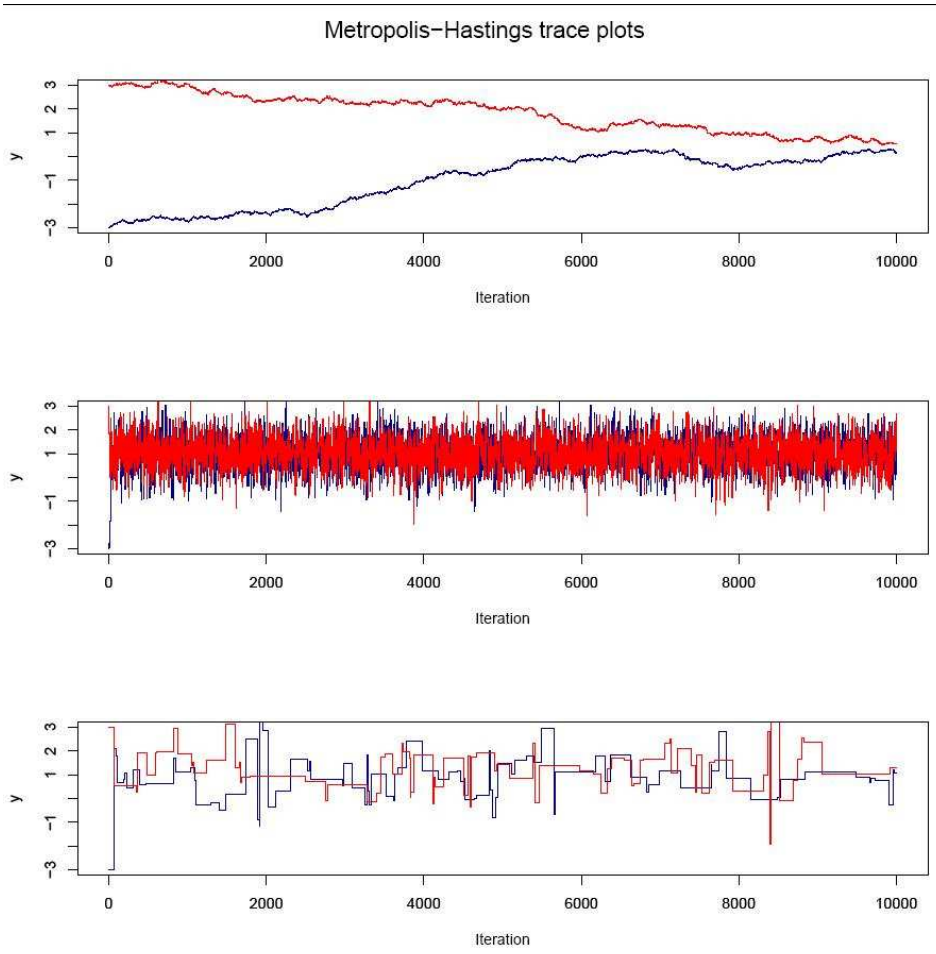

Figure 1.2: The effect of altering the variance of the proposal distribution on the mixing and convergence of a Metropolis algorithm. Each trace plot shows accepted points with a target distribution of $N(1,1)$. The proposal distribution is $N(\theta^{(i)}, \sigma^2)$, with σ^2 equal to 0.01, 1 and 100 from top to bottom. For each value, we run two independent chains (blue and red) with starting points at -3 and 3.

It is important to ensure that the MCMC algorithm has converged to the target distribution, and is mixing adequately. A simple and informal test is to run several independent MCMC chains with various starting points, in order to ensure they appear to sample from the same distribution after a burn-in period. This may identify potential multiple modes, which may not be discovered by running a single chain. More formally, Gelman and Rubin proposed a statistic to compare between-chain and within-

chain variance [121]. If convergence has occurred, these variances should be approximately the same. Geweke proposed a hypothesis test for convergence [122]. Rejection of the null hypothesis, that the first x percent and the final y percent of the MCMC samples have the same mean, indicates lack of convergence. While diagnostics such as these may indicate non-convergence, it is not possible to demonstrate conclusively that convergence has been attained in a finite MCMC sample [123].

1.5.2.5 Gibbs sampler

The Metropolis-Hastings algorithm may be conducted by updating the vector of parameters θ simultaneously, or repeating for individual elements, or groups of elements. Let

$$\theta_{-j}^{cur} = \{\theta_1^{(i+1)}, \ldots, \theta_{j-1}^{(i+1)}, \theta_{j+1}^{(i)}, \ldots, \theta_n^{(i)}\}$$

be the current vector of accepted parameters excluding θ_j. Then the acceptance probability in equation (1.5.5) for a proposed single element θ_j^* is

$$\alpha = \min\left(1, \frac{q(\theta_j^{(i)}|\theta_{-j}^{cur}, \theta_j^*)\pi(x|\theta_{-j}^{cur}, \theta_j^*)\pi(\theta_{-j}^{cur}, \theta_j^*)}{q(\theta_j^*|\theta_{-j}^{cur}, \theta_j^{(i)})\pi(x|\theta_{-j}^{cur}, \theta_j^{(i)})\pi(\theta_{-j}^{cur}, \theta_j^{(i)})}\right). \tag{1.5.6}$$

The full conditional distribution of a parameter θ_j is given as $\pi(\theta_j|\theta_{-j}, x)$, where θ_{-j} denotes the parameter vector θ without the jth component. If the full conditional distribution is of a form from which we may conveniently draw samples, then we can set the proposal distribution $q(\theta_j^*|\theta_{-j}^{cur}, \theta_j^{(i)})$ to the full conditional distribution for θ_j. Substituting into equation (1.5.6), we find that $\alpha = 1$ for all sampled θ_j^*. This process, a special case of the Metropolis-Hastings algorithm, is known as Gibbs sampling.

Gibbs Sampler

1. Set initial parameter values, $\theta^{(0)}$ and number of iterations, N.
2. For each element j, sample $\theta_j^{(i+1)}$ from full conditional distribution $\pi(\theta_j|\theta_{-j}^{cur}, x)$.
3. If $i < N$, go to step 2.

1.5.2.6 Reversible jump Markov chain Monte Carlo

The reversible jump Markov chain Monte Carlo (RJMCMC) algorithm is a generalisation of the Metropolis-Hastings algorithm to a multiple model setting [124]. Suppose we have data x, and assume that it has been generated under one of a set of competing models $m_1, \ldots, m_k \in \mathcal{M}$, each associated with parameters $\theta_1 \in \Theta_1, \ldots, \theta_k \in \Theta_k$. The RJMCMC algorithm samples points across the space $\mathcal{M} = \bigcup_{m=1}^{k} \{m\} \times \Theta_m$. Points are proposed according to the current model and parameter values $(m^{(i)}, \theta^{(i)})$, and transformation mechanisms to select a new model and to translate parameter values to a an 'equivalent' point in parameter space of the new model. Ideally, transformations are chosen such that the algorithm typically proposes points in a region of high posterior support, to encourage jumps between models.

The RJMCMC algorithm may be used to estimate within-model posterior densities, as well as the posterior probability for each model, $\pi(m|x)$. In this latter respect, it may be used as a tool for model comparison. In chapter 3, we discuss in detail the implementation of the RJMCMC algorithm, and its performance as a model comparison tool for transmission models.

1.5.2.7 Data augmentation

In many settings, we are faced with an intractable likelihood, $\pi(x|\theta)$. We aim to introduce missing data, z such that the likelihood $\pi(x, z|\theta)$ is tractable, where

$$\pi(x|\theta) = \int_z \pi(x, z|\theta) \mathrm{d}z.$$

By treating the missing data as a set of parameters to be estimated, we may use a data-augmented MCMC algorithm to sample from the posterior density of θ.

In the analysis of epidemics, unobserved transmission dynamics can result in an intractable likelihood. Augmenting the parameter space with infection times can result in a tractable likelihood, and as such, data augmentation methods provide a suitable framework to analyse epidemic models [52, 94, 125, 126].

To sample over missing data, a data-augmented MCMC algorithm is used, in which the parameter space is augmented with missing data z, treated as a set of parameters to be estimated. This vastly increases the parameter space, and often, the dimension of this object is uncertain.

The data-augmentation algorithm we present here is a special case of the reversible

jump MCMC algorithm, described in the previous section.

Data-augmented MCMC algorithm

1. Set initial parameter values, $\theta^{(0)}$, $z^{(0)}$, and number of iterations N.
2. Perform Metropolis-Hastings steps to update the set of parameters of interest, θ.
3. With probability density $g(z^*)$, propose to update the augmented data to z^*.
4. Accept this move with probability
$$\alpha = \min\left(1, \frac{\pi(x|z^*,\theta)\pi(z^*|\theta)\pi(\theta)g(z)}{\pi(x|z,\theta)\pi(z|\theta)\pi(\theta)g(z^*)}\right),$$
and set $z^{(i)} = z^*$, otherwise, set $z^{(i)} = z^{(i-1)}$.
5. If $i < N$, go to step 2.

Here, we assume that the augmented data are generated by an independence sampler; that is, the proposal function, $g(z^*|z,\theta) = g(z^*)$. The general form of the RJMCMC algorithm acceptance rate includes the Jacobian determinant of the transformation function. The independence sampler results in a Jacobian value equal to 1.

1.5.2.8 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) is a likelihood-free method of parameter estimation, in which data are repeatedly simulated under proposed parameter values, with the aim of eventually simulating data which approximate the observed data to within some defined threshold. This approach is of most use in cases where the likelihood is complex or intractable, but data may be simulated relatively easily.

The simplest approach is the ABC rejection sampler algorithm, in which parameters θ^* are successively drawn from their prior distribution, and data are simulated based on θ^*. If the simulated dataset x^* is similar to the observed data x, that is, $\delta(x, x^*) < \epsilon$ for some distance metric $\delta(\cdot,\cdot)$ and threshold ϵ, the sampled parameters θ^* are accepted [127]. Unless the prior can be chosen to be similar to the posterior distribution, which

is often not possible, this process is highly inefficient. This may be improved upon by adopting a MCMC-based sampling algorithm, in which the acceptance rate depends on both the similarity of the simulated dataset and the previous parameter sample [128].

1.5.2.9 Sequential Monte Carlo

Sisson et al. described an ABC algorithm based on sequential Monte Carlo (SMC) methods [129]. In this approach, a number of 'particles', $\theta_1^{(1)}, \ldots, \theta_1^{(N)}$, are initially drawn from the prior distribution $\pi(\theta)$. At each step i of the algorithm, the particles are updated, such that θ_i represents a sample from an intermediate distribution $\pi(\theta|\delta(x, x^*) \leq \epsilon_i)$, where $\epsilon_1 > \cdots > \epsilon_T > 0$ is a sequence of decreasing thresholds to ensure data simulated from particles come closer to approximating the observed dataset. A version of the algorithm runs as follows:

ABC SMC

1. Set initial tolerances, $\epsilon_0, \ldots, \epsilon_T$, and set tolerance level indicator $t = 0$.

2. Set particle indicator $i = 1$.

3. If $t = 0$, sample $\theta^{**} \sim \pi(\theta)$, otherwise draw θ^* from the $(t-1)$th sample, $\{\theta_{t-1}^{(j)}\}_{j=1}^N$, according to weights $\{w_{t-1}^{(j)}\}_{j=1}^N$. Generate θ^{**} using a perturbation kernel, $K_t(\theta|\theta^*)$.

4. Simulate data x^* under the proposed particle θ^{**}. If $\delta(x, x^*) \geq \epsilon_t$, then go to step 3.

5. Set $\theta_t^{(i)} = \theta^{**}$, and

$$w_t^{(i)} = \begin{cases} 1 & t = 0, \\ \frac{\pi(\theta_t^{(i)})}{\sum_{j=0}^N w_{t-1}^{(j)} K_t(\theta_{t-1}^{(j)}|\theta_t^{(i)})} & t > 0. \end{cases}$$

   If $i < N$, set $i = i + 1$, then go to step 3.

6. Normalise weights. If $t < T$, set $t = t + 1$ and go to step 2.

Clearly, the performance of SMC relies on the specification of the distance metric δ, as well as the perturbation kernel and tolerance levels. For all but the simplest of cases, it is difficult, if not impossible, to specify an adequate statistic with which to measure the closeness, or similarity, of two datasets. Without careful calibration, the algorithm can be inefficient, and can lead to poor estimates.

1.5.2.10 Posterior predictive distribution

Given data x, we fit a model and evaluate the posterior density for a set of parameters θ, $\pi(\theta|x)$. Similarly, we can define the posterior predictive distribution of a hypothetical future dataset $\tilde{X}$ as $\pi(\tilde{X}|x)$, defined as

$$\pi(\tilde{X}|x, A(x)) = \int \pi(\tilde{X}|\theta, A(x))\pi(\theta|x)\mathrm{d}\theta, \tag{1.5.7}$$

where $A(x)$ are 'auxiliary statistics' which are matched when sampling replicated data, so that $A(\tilde{X}) = A(x)$ [130]. This may involve matching sample size, study duration, etc.

The posterior predictive distribution can be used as a measure of goodness-of-fit in a Bayesian setting. The posterior distribution of $\tilde{X}$ will generally be complex, and as such, it is useful to derive a summary statistic $T(X)$, which may be used to assess the similarity of datasets. By repeatedly sampling datasets $\tilde{X}$ from the posterior predictive distribution, we may approximate the distribution of $T(\cdot)$. The posterior predictive p-value can then be calculated by comparing the realised value, $T(X)$ to this distribution [131], and is defined as

$$PPV = P(T(\tilde{X}) < T(x)|\theta, x).$$

Care should be taken in the choice of T, in order to ensure it adequately summarises any discrepancies between datasets. For instance, suppose we wish to compare two sets of hospital surveillance data, comprising positive and negative results. If we were to consider total number of positive and negative results as measures of similarity, this would be insufficient, since the same results could arise from very different transmission dynamics. This could potentially generate misleading results.

1.6 Bayesian model choice

Given a set of possible models $m_1, \ldots, m_k$, it is of interest to evaluate the relative support for each model given by the data x, and to determine a 'best' model. Many meth-

ods exist to compare models in a Bayesian setting, but the performance and interpretation of such methods are debated. Of fundamental interest is the posterior model probability,

$$\pi(m|x) \propto \int_{\Theta_m} \pi(x|\theta_m, m)\pi(\theta_m|m)\mathrm{d}\theta_m \pi(m),$$

where $\pi(x|\theta_m, m)$ is the likelihood of the data under model m, $\pi(\theta_m|m)$ is the prior density of parameters θ_m for model m, and $\pi(m)$ is the prior probability of model m. As described in section 1.5.2.6, the RJMCMC algorithm may be used to estimate posterior model probabilities by sampling over the joint parameter space of all models. This approach can, however, be difficult to implement, and requires careful calibration.

One of the most well known Bayesian model comparison tools is the Bayes factor. This evaluates the evidence in favour of one model over another, conditional on one of the models being true. Since its first publication in 1935 by Jeffreys [132] (then described as a 'significance test'), it has been widely used for Bayesian hypothesis testing and model comparison, despite often being difficult to calculate and highly dependent on the choice of model-specific prior distributions [133–135].

The Bayes factor of models l and m is the ratio of marginal likelihoods,

$$BF(l,m) = \frac{\pi(x|l)}{\pi(x|m)} = \frac{\int \pi(x|\theta_l, l)\pi(\theta_l|l)\mathrm{d}\theta_l}{\int \pi(x|\theta_m, m)\pi(\theta_m|m)\mathrm{d}\theta_m}. \tag{1.6.1}$$

The greater the value of the Bayes factor $BF(l,m)$, the stronger the evidence in favour of model l — an interpretation of the magnitude of the Bayes factor was provided by Kass and Raftery [133], and is shown in table 1.1. Clearly, very small values correspond to evidence in favour of model m, since $BF(l,m) = 1/BF(m,l)$.

Bayes factor interpretation	
$BF(l,m)$	Evidence in favour of model l
1–3	Not worth more than a bare mention
3–20	Positive
20–150	Strong
> 150	Very strong

Table 1.1: An interpretation of Bayes factors, as given by Kass and Raftery [133].

The Bayes factor can also be rewritten as the ratio of the posterior model odds and the prior model odds

$$BF(l,m) = \frac{\pi(l|x)}{\pi(m|x)} \frac{\pi(m)}{\pi(l)}, \tag{1.6.2}$$

and as such, the Bayes factor represents the extent to which our beliefs about the relative probabilities of the models have altered, having observed the data x. Often, little is known about the model prior to analysis, and as such, models are assigned equal prior probability. In this case, the Bayes factor is simply equal to the posterior model odds.

The Bayes factor is often difficult to evaluate due to the intractability of the marginal likelihood $\pi(x|m)$. We discuss this in further detail in section 3.2.2.

Information criteria are scores which measure the relative adequacy of a given model. This involves a trade-off between parsimony and relative fit, the aim being to select a simple model which fits well, compared to alternative models. The deviance information criterion (DIC) is a generalised model comparison tool which can be used in a Bayesian setting, and is analogous to Akaike's Information Criterion (AIC) [136]. Model complexity is represented by the number of parameters in the AIC, but this concept is poorly defined in a Bayesian setting. Suppose we have data x, and a model defined by a likelihood function $\pi(x|\theta)$. Then the deviance is defined as

$$D_x(\theta) = -2\log \pi(x|\theta),$$

and the DIC is defined as

$$\text{DIC} = D(\tilde{\theta}) + 2p_D, \tag{1.6.3}$$

where p_D is the 'effective number of parameters', and $\tilde{\theta}$ is a point estimate of θ, such as the posterior mean. This represents a trade-off between model adequacy and complexity. Models with a lower DIC value are considered to be a better fit to the data, although it is not possible to quantify the improvement between models with this measure.

For models with missing data, the likelihood, and therefore the deviance, is not available in closed form. This means the DIC in this form cannot be implemented. Celeux et al. proposed a range of alternative formulations of DIC to account for latent data [137]. We discuss the DIC, and a missing data variant of it, in greater detail in section 3.5.

1.7 Epidemic models for HCAIs

Modelling the spread of pathogens in a hospital setting gives rise to a specific set of challenges. Firstly, the population is typically very small — at any given time, the population is limited to the number of beds in the hospital ward. This means that deterministic models are inappropriate, since random fluctuations in prevalence are likely to play an extremely important role in transmission dynamics at a ward level

[27, 94, 95]. Secondly, the ward population is dynamic; individuals regularly enter and exit the population. MRSA outbreaks in hospitals are characterised by one or more introductions of the pathogen to the ward from either another ward, or the community. Transmission may then occur from this source to susceptible individuals on the ward, and an outbreak is at an end when all positive patients have either been removed from the population (death or discharge), or have received antibiotic treatment so that their bacterial carriage is either removed, or reduced to levels at which transmission may not occur. Thirdly, carriage of some nosocomial pathogens, including MRSA, is often asymptomatic. This means that we cannot be certain of a patient's status without screening tests. Typically we do not know an individual's disease state on admission to the ward. Additionally, screening tests do not provide perfect observations, and we must allow for the possibility of false negative results [138]. Finally, it is widely accepted that MRSA is spread indirectly between individuals, via contact with healthcare workers. We consider an MRSA transmission event in this setting to be the temporary colonisation of a HCW, via contact with an MRSA-positive individual, and a subsequent transmission of this bacterial colony to a susceptible individual. The probability of such a series of events occurring depends on the type of HCW-patient contact, the occurrence of hand washing between patient contacts, and the susceptibility of the patient at risk of colonisation, according to risk factors such as those described in earlier, in section 1.2.2.2. This indirect transmission mechanism requires consideration of HCW-patient contact patterns, rather than mixing of patients themselves. While we might assume homogeneous mixing in a standard SIR epidemic model, the equivalent in this setting is the assumption that there is an equal probability of a HCW contacting any two patients in succession. If two patients are at no point visited by the same HCW, there is a low chance of transmission occurring between them.

1.7.1 Early approaches

The earliest HCAI transmission models were simulation-based and deterministic, such as that created by Sébille et al. in 1997 [139]. This model allowed patients and HCWs to be treated as two populations in a ward, and was created to assess various infection control measures. This study supported the connection between colonised HCWs and patient acquisition, and the reduction in transmission due to increased hand hygiene compliance. D'Agata et al. similarly used a deterministic compartmental model to investigate transmission of vancomycin-resistant Enterococci (VRE) in 2002 [140]. While

a deterministic model such as this can sufficiently assess the impact of interventions for a large population, random fluctuations in prevalence and population play a much more important role on a ward level, and as such, a stochastic model is a more appropriate way of modelling such a setting than a deterministic structure [27, 94, 141].

Recognising the limitations of deterministic modelling for a HCAI dynamics, Cooper et al. conducted a simulation study, using a stochastic transmission model, to investigate the impact of hand hygiene compliance and surveillance on the transmission of nosocomial pathogen transmission [27]. Similarly, Austin et al. investigated VRE transmission using a simulation study [142]. Both studies emphasised the importance of stochasticity in modelling transmission in a small population.

1.7.2 Markov models

A Markov model for the prevalence of nosocomial infections was introduced by Pelupessy et al. in 2002 [143]. This was used to analyse longitudinal prevalence data for VRE and *Pseudomonas aeruginosa* from hospital ICUs. In this analysis, cases are assumed to be perfectly observed, and the ward is a fixed size, where discharged patients are immediately replaced by admitted patients. Patients are all screened simultaneously at regular intervals. If A is a transition rate matrix, where $A_{i,j}$ gives the rate of moving from population level i to j, then the probability of moving from i to j colonised patients in time δt is given by the (i, j)th entry of the matrix exponential $\exp(A\delta t)$. Now according to this model, the probability of a set of observed prevalences $x_1, \ldots, x_n$ at times $t_1, \ldots, t_n$ is given by the product

$$\prod_{i=1}^{n-1} \exp(A(t_{i+1} - t_i))_{x_i, x_{i+1}}.$$

Estimates for the parameters which govern the matrix A can then be derived through the optimisation of this function.

This model was extended by Cooper and Lipsitch in 2004; a structured hidden Markov model was introduced to describe unobserved transmission dynamics based on observed monthly nosocomial incidence data [144]. Asymptomatic carriage means that prevalence is likely to be higher than the number of observed cases presenting clinical symptoms. The authors define a Poisson observation model of the underlying transmission process, so that

$$P(Y(t) = y | X(t) = x) = \frac{e^{-\lambda x}(\lambda x)^y}{y!},$$

where $X(t)$ is the true number of colonised patients at time t, and $Y(t)$ is the number of observed infected cases. Population counts are bounded by the ward size, n_w. Likelihood optimisation is used to derive parameter estimates. As the state space increases with population size n_w, computation can become slow. Problems can arise with the calculation of the MLE and confidence intervals, particularly with smaller amounts of data.

This model was adapted for use in a Bayesian framework by McBryde et al. in 2007 [145]. Weekly VRE prevalence data taken from ICUs are assumed to have been imperfectly observed, and a binomial observation model is placed on the unobserved true process, based on a fixed detection rate (sensitivity). The authors defined a transmission rate dependent on the number of carriers present in the ward, plus a constant rate of 'sporadic' acquisition. A Markov chain Monte Carlo algorithm is used to sample parameters from the posterior distribution. At each iteration, hidden states $X(t)$ are updated using a Gibbs sampler, based on the conditional probability

$$P[X(t)|X^{-}(t), Y] \propto P[Y(t)|X(t)]P[X(t+1)|X(t)]P[X(t)|X(t-1)],$$

where $X^{-}(t)$ denotes the set of true states, excluding that at observation time t.

McBryde et al. proposed a bivariate system, describing the changing number of colonised patients and HCWs over time [146]. This was further adapted by Drovandi and Pettitt [147], who considered a trivariate system by incorporating incidence data, building on the model by Cooper and Lipsitch [144]. This can be simplified to a bivariate system by assuming the HCW prevalence to be at an equilibrium. Maximum likelihood techniques were used to obtain transmission parameter estimates. The authors reported problems in estimating parameters in the trivariate system, potentially due to over-parametrisation. The dimension of the state space for such a system is $(n+1)^2(m+1)$, where n is the ward capacity, and m is the (arbitrarily chosen) maximum incidence.

The authors conducted a similar analysis using approximate Bayesian computation (ABC) methods in 2012 [148]. Computing the likelihood of these joint processes is computationally expensive, more so than simulating these dynamics, making ABC methods an attractive alternative to likelihood optimisation in this setting. However, as mentioned earlier, ABC methods are very dependent on the distance metrics chosen to measure similarity of datasets, and inappropriate threshold limits can lead to misleading results.

Aggregated daily or weekly prevalence counts, as often used in Markov models such as those described in this section, are clearly less informative than individual-level data,

but are more readily obtained from hospitals. Each of the methods described above requires the estimation of a discharge rate for colonised individuals, and the assumption that the ward is full to capacity at all times. The approach is unsuitable for large state spaces, due to the requirement to calculate the matrix exponential in the likelihood.

A discrete-time Markov algorithm was introduced by Bootsma et al. to assess the role of cephalosporin-resistant Enterobacteriaceae (CRE) transmission routes for hospital pathogens [149]. This algorithm considers patients to have a status of 'susceptible', 'colonised' or 'unknown'. A probability of carriage is calculated for each patient of unknown status, based on screening results and observed prevalence. Maximum likelihood estimates are calculated for both endogenous and exogenous transmission rates. The authors state that running the algorithm becomes problematic when the number of patients with an unknown status is larger than 10. By assuming perfect sensitivity, this number can be kept fairly low.

1.7.3 Data augmentation methods

MCMC methods lend themselves well to dealing with missing data, a common issue in the analysis of epidemics. Some of the first Bayesian analysis for epidemic modelling was published by Gibson and Renshaw in 1998 [150]. This study analysed simulated data using an SIR model, and accounted for hidden events using RJMCMC methods. The study noted that this methodology would be applicable to real data, such as *Clostridium difficile* hospital surveillance data. O'Neill and Roberts then introduced an MCMC approach to estimate transmission parameters for a stochastic SIR epidemic model [94]. Here, unobserved infection times were inferred using a data-augmented MCMC algorithm. This method was tested using simulated data and then applied to smallpox prevalence data collected from Nigeria.

Following this work, other similar approaches have been employed to tackle the problem of partially-observed data in the analysis of epidemics. Forrester et al. applied this method to MRSA transmission in a hospital setting, using individual-level data, and incorporating admission and discharge events, as well as importation and imperfect test sensitivity [59]. This approach allows the investigation of heterogeneous transmission, which was used to measure the difference in the transmission rate from isolated and unisolated patients. Cooper et al. conducted an analysis of VRE transmission in a hematology unit, comparing transmission rates under differing antibiotic usage policies [126]. Kypraios et al. employed this method to determine the effect of isolation

precautions on MRSA transmission rates in ICUs [52].

1.7.4 Utilising WGS data for transmission analysis

As yet, there have been no studies combining epidemiological data and genetic data for the analysis of nosocomial transmission routes of which we are aware. There are, however, some studies which have attempted to achieve this in different settings. Cottam et al. utilised sequence data to create a set of plausible transmission trees, and then calculate a likelihood of each of these trees based on the infectiousness of each individual [151]. This approach was used to investigate the spread of foot-and-mouth disease between 20 farms in the UK. This approach is likely to become computationally expensive for larger datasets, and does not integrate genetic and epidemiological data, instead using sequence data to produce a set of possible transmission networks, for which a likelihood is provided, based on epidemiological data. Jombart et al. described a network optimisation approach, in which edges represent infection routes between hosts (nodes), and are weighted by the genetic distance between isolates taken from each individual [152]. This method assumes the minimal weight network is the best reconstruction, with no indication of uncertainty. Furthermore, transmission times are restricted by the collection times of isolates: individuals may not be infected prior to observation. Ypma et al. provided a framework to investigate the transmission of avian influenza in the Netherlands, using spatial, temporal and genetic data [65]. The model assumes these data are independent, and the likelihood is simply the product of contributions for each data type.

These methods do not take into account multiple sequences for each individual; as such, within-host genetic diversity is ignored. These methods have limitations which would make the analysis of MRSA transmission in hospitals impossible. Firstly, transmission networks typically have multiple origins, representing the importation of an MRSA carrier into the hospital or ward. Secondly, asymptomatic carriage and imperfect tests mean that colonisation times are uncertain, and it is important to account for this if attempting to estimate a transmission source.

1.8 Conclusion

In this chapter, we have discussed the importance and impact of healthcare-associated infections, in particular, MRSA. We described various interventions to curtail the spread

of the pathogen in hospitals. In the next section, we introduced epidemic modelling, and discussed the role of statistical analysis in the investigation of transmission dynamics. We outlined both frequentist and Bayesian methods of parameter inference, describing in particular methods to deal with partially-observed/missing data, an issue commonly encountered in epidemic modelling. Finally, we discussed the existing literature on HCAI modelling, which sought to assess MRSA transmission dynamics, and the success of interventions to reduce this. We described some of the existing literature in which genetic data is utilised to reconstruct transmission networks.

1.9 Aims and structure of the thesis

In this thesis, we aim to do the following:

1. Investigate the transmission dynamics of MRSA in hospital general wards and evaluate the effectiveness of infection control measures, using a stochastic modelling approach.
2. Systematically review the performance of different Bayesian model comparison techniques, applied to HCAI transmission models.
3. Develop new methods to incorporate WGS data into the analysis of HCAI outbreaks, in order to investigate routes of transmission.

In chapter 2, data augmentation methods are applied to individual-level patient data, in order to model MRSA transmission in several general medical wards, accounting for imperfect observations. Most studies of MRSA transmission have been set in intensive care units (ICUs), and little is known about transmission dynamics in this setting, despite the potential role of general medical wards as an MRSA reservoir to the rest of the hospital. As discussed in section 1.2.3, the impact of interventions such as isolation are of great interest in terms of infection control and cost-effectiveness evaluations of hospital policy. A discrete-time data-augmented MCMC approach is used to estimate the effectiveness of using patient isolation and decolonisation treatment to reduce the spread of the pathogen. The performance of the MCMC analysis is compared with a pseudolikelihood approach, which could potentially offer a less computationally-intensive method to derive parameter estimates. Finally, the incorporation of an additional dataset of clinically informed observations is investigated.

In chapter 3, we explore Bayesian model selection methods, with the aim of determining plausible underlying transmission models, given a set of individual-level patient data. In section 1.6, Bayesian model choice was introduced, and some of the challenges faced in this field were briefly described. There are many approaches to model choice in a Bayesian framework, each with advantages and disadvantages. We review some of the existing literature on Bayesian model selection, before conducting a systematic simulation study on the performance of two of the more widely-used methods, RJMCMC, and the DIC. We describe the conditions necessary for these approaches to give reasonable results for HCAI transmission scenarios, and discuss limitations of the two methods. There are few studies which systematically study the performance of Bayesian model choice methods for a particular setting. Since it is of great interest to compare different transmission models, this work provides an insight into situations where RJMCMC and DIC can be useful tools. These methods are applied to compare transmission models, using the MRSA carriage data described in the second chapter.

In chapter 4, we consider the incorporation of whole genome sequence data into a hospital transmission model. A dataset collected from two ICUs in Thailand is used, in which sequence data were collected from MRSA carriers. In section 1.7.4, existing methods to integrate epidemiology with genetic data were reviewed. Such approaches have many limitations and restrictions which would prevent the analysis of MRSA transmission in hospitals. In chapter 4 we describe new methods which address some of these limitations in order to analyse transmission in the Thai ICUs. We firstly investigate whether a difference in transmission may be detected between genetically different groups. We then attempt to reconstruct the unobserved transmission routes in the ICUs, examining two alternative models for genetic diversity. The models are tested with a series of simulated datasets, before being applied to the Thai data. The increasing availability of WGS data has created a demand for statistical methods to exploit this additional information to gain a greater insight in the dynamics of communicable pathogens, and we present here novel methods to do this.

Chapter 5 concludes the thesis with a summary of findings, and the contributions to the study of nosocomial pathogen transmission made by this thesis are described.

CHAPTER 2

The effectiveness of patient isolation and decolonisation treatment in reducing MRSA transmission

This chapter forms the basis of a paper titled *Estimating the effectiveness of isolation and decolonisation measures in reducing MRSA transmission in hospital general wards*, by C. J. Worby, D. Jeyaratnam, J. V. Robotham, T. Kypraios, P. D. O'Neill, D. De Angelis, G. French and B. S. Cooper, to appear in the American Journal of Epidemiology.

2.1 Introduction

Patient isolation plays a central role in many local and national infection control guidelines and policies [153–155], however, there is much debate concerning its effectiveness at reducing MRSA transmission in healthcare facilities. Existing studies have investigated the effectiveness of isolation in ICUs [52, 57, 156], and veterans affairs hospitals [157], with mixed results. The study of MRSA transmission dynamics and evaluation of intervention policies in general medical wards has been given considerably less attention. In this chapter, data from ten general wards in a London hospital are analysed with the aim of estimating the effectiveness of isolation and decolonization measures in reducing MRSA transmission rates.

In section 2.2, we begin by discussing the role of patient isolation and decolonisation

treatment in hospital policies, and the debate surrounding the use of isolation as a component of infection control strategies. We also discuss the importance of assessing MRSA transmission dynamics in a general ward setting, particularly to understand the role of interventions. The dataset used in this study is described in section 2.3, as well as the infection control policies used by the hospital. In section 2.4, we introduce stochastic models to describe transmission dynamics in this setting. A data-augmented MCMC algorithm is described which accounts for unobserved colonisation times, and is used to derive model parameter estimates, as well as a measure of isolation and decolonisation effectiveness. We also consider an alternative, less computationally-intensive analysis of the data, using a pseudolikelihood approximation. The results of this faster method are compared with those derived from the MCMC algorithm in section 2.5. We discuss results and the implication of our findings in a wider context. Finally, in section 2.7, we describe an additional dataset collected from the hospital, comprising additional swabs from various body sites taken from individuals considered high risk of MRSA carriage. These are not included in the first analysis for reasons of bias, but we develop methods which allow these data to be incorporated to the analysis, and discuss the loss of information due to their exclusion. We conclude the chapter with a discussion of our findings.

2.2 Background

2.2.1 Patient isolation

The isolation of MRSA carriers in hospitals is employed as a means to protect susceptible patients from the risk of colonisation or infection via cross-transmission. Isolation measures take the form of single room confinement, patient cohorting (grouping carriers in one part of a ward), or simply enhanced contact precautions on the ward (the use of gowns and gloves when dealing with known carriers). The level of isolation employed may depend on the availability of resources and the risk of transmission. It has been recommended that carriers are isolated in a single side room if resources permit, or in a cohort with other positive patients [21, 53]. Isolation is typically employed as part of an infection control package, and the colonised patient may additionally receive further interventions, such as decolonisation therapy.

Patient isolation is costly, and hospitals have a limited capacity for single-room isolation. The question of if and when to employ isolation for suspected of confirmed MRSA

positive patients is controversial [54]; there have been few studies which quantify the effectiveness of isolation (or lack thereof) [158]. The systematic review of isolation usage conducted by Cooper et al. [144] in 2004 concluded that there was a lack of well designed studies in this area. In addition, there are concerns that isolation measures may actually have a negative impact on patient welfare. Isolated patients feel unhappier and more neglected than unisolated patients [159, 160], and may be visited less often by HCWs [161–163], potentially resulting in a lower quality of care.

2.2.2 Previous studies

Questionnaire-based studies at a hospital level have been conducted in order to determine whether isolation precautions are associated with a lower MRSA prevalence [164, 165], and while these have indicated a lower transmission rate amongst hospitals which isolate MRSA positive patients, the number of confounding issues and potential biases mean that these provide little evidence towards the effectiveness of isolation.

Jernigan et al. conducted a small study in which typing data were used to identify probable sources of colonisation, in order to determine the relative transmission rates from isolated and non-isolated MRSA carriers [166]. They found a significantly lower rate of transmission for those under contact isolation; however, this study relies on subjective assessment of transmission routes, which may bias results.

In 2005, Cepeda and et al. undertook a prospective trial during which positive patients were not isolated for a period of six months in two intensive care units in London [57]. This period was then compared to a control phase in which isolation in a side room or cohort was used. Standard precautions were maintained throughout. Acquisition rates were found to be similar in both periods, and there was no evidence to suggest increased transmission during the period with no isolation. Potentially, the long turnaround time for screening results (3 days) may have reduced the apparent effectiveness of isolation, since this delayed its implementation.

Huskins et al. conducted a cluster-randomised trial over six months to assess the effect of expanded barrier precautions (increased usage of gowns and gloves, improved hand hygiene) in reducing MRSA and VRE incidence in ICUs [156]. No reduction was observed, although the required compliance to intervention measures was not attained. In addition, the study attracted much criticism for the turnaround time for swab results — swabs were sent for processing offsite. Jain et al. conducted a trial with similar aims, in veterans affairs healthcare facilities [157]. They assessed the impact of an 'MRSA

bundle', which involved universal surveillance for nasal MRSA carriage and increased contact precautions and hand hygiene when dealing with known carriers. The study, conducted over a period of 32 months, found a reduction in MRSA infections of 62% in ICUs and 45% in non-ICU wards associated with the introduction of the bundle. However, it was later demonstrated that the transmission prevention component of the bundle was likely to have played only a small part in the reduction [167].

There have also been model-based analyses of isolation effectiveness using regular carriage surveillance data. Cooper et al. developed a dynamic transmission model to theoretically evaluate isolation, and demonstrated the importance of isolation capacity and timing to the success or failure of isolation in reducing transmission [144]. Forrester et al. described an interval-censored approach, in which the number of positive patients in each interval was modelled as a binomial distribution, dependent on patient numbers in the previous interval [58]. This approach assumes colonisations in each interval are independent. The authors used ICU patient surveillance data, and found weak evidence to suggest the transmission rate from colonised and isolated patients was less than that of unisolated carriers. A later study on the same data, using a data-augmented MCMC approach, confirmed the lower transmission rate associated with isolated MRSA positive patients [59].

In 2010, Kypraios et al. conducted another model-based evaluation of barrier precautions, using data collected in several ICUs in Boston [52]. They found some evidence to support the use of isolation; a best estimate of 28% reduction in transmission was given, with a relatively high degree of uncertainty.

2.2.3 Aims

The majority of analyses of MRSA transmission in hospitals use data collected in ICUs, therefore little is known about transmission dynamics in general medical wards. General wards have a highly dynamic population with frequent readmissions and ward transfers. While individuals in general wards typically have a lower antibiotic consumption and are less susceptible to infection than those in intensive care, many more patient days are spent in general wards, making them potentially important reservoirs for MRSA and locations for MRSA transmission within the hospital.

We present an analysis of individual-level MRSA carriage data collected from a selection of hospital general wards in 2007–08. In this study, our primary aim was to estimate the combined effect of isolation and decolonisation treatment in reducing the

Ward characteristics

Ward no.	Specialty	Location	Patient episodes	Episode length median days (IQR) [mean]
1	Surgery (plastics)	St. Thomas'	1808	2.8 (1.4, 6.1) [5.5]
2	Elderly care	St. Thomas'	644	11.0 (6.2, 21.7) [16.5]
3	Surgery (urology)	Guy's	2249	2.2 (1.3, 4.0) [3.5]
4	Surgery (ear, nose & throat)	Guy's	2825	2.0 (1.1,3.3) [4.0]
5	Surgery (cardiothoracic)	Guy's	1619	4.3 (2.1, 7.1) [5.5]
6	Elderly care	St. Thomas'	641	10.2 (5.9, 21.8) [16.3]
7	Surgery (vascular)	St. Thomas'	1319	4.0 (1.9, 8.8) [7.7]
8	Surgery (gastrointestinal)	St. Thomas'	1293	3.9 (1.7, 8.6) [6.9]
9	Oncology	Guy's	797	6.0 (3.4, 14.6) [11.0]
10	Oncology	Guy's	840	5.4 (2.3, 12.0) [9.3]

Table 2.1: Characteristics of the wards included in the study. Mean and median study length is given, along with the interquartile range (IQR).

transmission rate of MRSA in hospital general wards, and to assess the importance of different MRSA transmission routes. In order to do this, we constructed stochastic epidemic models to describe the transmission dynamics within each ward, and fitted these to patient data. By doing so, we could assess the importance of colonisation pressure in determining MRSA transmission, and estimate how much transmission was attributable to a constant background effect unrelated to colonisation pressure, due, for example, to long-term staff carriers and environmental contamination. In addition, we aimed to compare the effectiveness of single room isolation to isolation measures on the open ward (both in combination with decolonisation treatment). Estimates for the probability of being colonised on admission and the sensitivity of the screening test were also derived.

2.3 Data

We used data collected between January 2006 and April 2007 at Guy's and St. Thomas' hospital (GST), a teaching hospital on two sites in London, as part of a prospective clus-

ter randomised crossover trial to determine whether a policy of rapid screening (with a polymerase chain reaction (PCR) test) for MRSA could reduce the rate of acquisition compared with a policy of conventional culture screening (both combined with isolation and decolonisation treatment for positive patients). This study found no evidence to demonstrate a reduction in transmission associated with the use of rapid screening [50]. Our analysis has fundamentally different aims to the original study, and the methods used are quite different. Furthermore, the data we used span a longer time period than the data considered previously.

Data were collected across ten hospital general wards, comprising of surgery, elderly care and oncology wards. Characteristics of the ten study wards are given in table 2.1. All patients were culture screened within 48 hours of admission where possible, and most patients were also culture screened on discharge. Table 2.2 provides a summary of patient admissions and culture swab results for all ten study wards. Full details of data collection, microbiological methods and ethical approval were reported by Jeyaratnam et al. [50].

Patients considered high risk for MRSA carriage on admission were isolated where possible prior to the admission swab result. A decision to implement this 'pre-emptive isolation' was based on a previous MRSA positive swab or the presence of one or more risk factors for MRSA carriage (living in a nursing or residential home, an inpatient stay during the previous year, a direct transfer from another hospital, from abroad or from a high risk area within GST). Patients found to be MRSA positive by the admission screen (by culture or PCR) were also isolated. A single side room was used if available and appropriate; otherwise the positive patients were nursed on the open ward with standard contact precautions (staff wore disposable gowns and gloves). The isolation policies and practices were strictly enforced. When more than one MRSA positive patient had to be on the open ward, they were placed together in a separate bay where possible (patient cohorting). Decolonisation treatment using chlorhexidine for the skin, povidone iodine or silver sulphadiazine for colonised wounds, and, for sensitive strains, mupirocin nasal ointment, was initiated for all patients found to be MRSA positive.

We made use of the full set of culture swabs taken over the 16 month period, in contrast to the original study by Jeyaratnam et al., where only data collected from two study periods of five months within this time frame were used [50].

2.4 Methods

2.4.1 Transmission model

Primarily, our interest lies in the dynamics of transmission. We supposed that each patient present in the ward is in either a 'susceptible' (MRSA negative) or a 'colonised' (MRSA positive) state at any given time. This latter state includes patients with asymptomatic carriage as well as those with MRSA infection. For any given susceptible patient, define $q(t)$ to be the transmission rate at which they may become colonised, at time t. This rate is dependent on the colonisation pressure in the ward at the time, which we here consider to be the number of MRSA positive patients present in the ward. Each susceptible patient is regarded as independent, and has the same probabil-

Patient screening statistics	
Number of days	452
Number of unique patients	10845
Number of patient episodes	14035
Number of patient days	94747
Mean length of stay (days)	6.8
Number of patients not screened on admission	649
(of total admissions)	(4.62%)
No. positive swabs on admission	649
(of those swabbed)	(4.84%)
Discharge swab invalid/missing	2687
(of those with admission swab)	(20.1%)
Stay <48hrs	2810
(of those with admission swab)	(21.0%)
Number of eligible observation pairs	8595
(of total episodes)	(61.2%)
Number of observed MRSA acquisitions	265
(of eligible observation pairs)	(3.08%)

Table 2.2: Summary statistics on the culture swab based screening data collected from the study wards. 'Eligible' pairs are defined as those in which both swabs were taken at the correct time and the patient stayed for 48 hours or more. This is for the purposes of counting acquisitions based purely on the observational data. Our analysis used all swab results to derive estimates.

ity of acquisition at a given time t. We considered three different transmission models, which differ by the formulation of the transmission rate. These are given below.

$$q_1(t) = a_0 + a_1 C(t) \quad (m=1)$$

$$q_2(t) = a_0 + a_1 C_N(t) + a_2 C_I(t) \quad (m=2)$$

$$q_3(t) = a_0 + a_1 C_N(t) + a_2 C_W(t) + a_3 C_S(t) \quad (m=3)$$

Constructing different models allows us to perform different analyses of MRSA transmission dynamics. Model 1 has a transmission rate dependent on the number of positive patients ($C(t)$), and a background transmission effect (a_0), which may arise from long-term staff carriers, persistent environmental contamination, or the introduction of the pathogen from elsewhere in the hospital. This model assumes all MRSA positive patients to be equally transmissible.

More generally, we have a multiple population model, where g different groups of colonized patients exist, and are associated with different transmission rates. For example, one could partition positive patients by age group, or by the type of antibiotic administered (if any). We assumed that individuals in each group are homogeneous in terms of transmissibility; that is, each patient in each group has the same potential to transmit the pathogen to a susceptible individual. Each susceptible patient is under colonisation pressure from each group independently. Since interest lies in the effect of isolation, model 2 is defined to be a two-population model, in which the transmission rate experienced by a susceptible patient may differ from isolated ($C_I(t)$) and non-isolated ($C_N(t)$) colonised patients ($C_I(t) + C_N(t) = C(t)$). It is assumed that all isolated colonised patients are equally transmissible. Similarly, all unisolated colonised patients are assumed to have an equal potential to transmit.

Model 3 allows the effect of positive patients in a side room ($C_S(t)$), and those receiving isolation precautions on the open ward ($C_W(t)$), to be considered separately. We defined 'open ward isolation' to cover the implementation of barrier precautions for patients either in a regular ward bay, or part of a patient cohort.

We considered a discrete time model, and assumed events occur within daily intervals, since data are not available at any greater resolution. On a given day t, susceptible patients were assumed to be under colonisation pressure from those colonised on day $t-1$ or earlier, as well as importations on day t. Colonisation pressure remains constant for the duration of day t. Any patients becoming colonised on day t were assumed to remain colonised for the duration of their stay, and contribute to colonisation pressure from day $t+1$ onwards. The probability of acquisition for any given susceptible patient

on day d in model i is then $1 - e^{-q_i(d)}$.

The episodes of any patients who are discharged and then later readmitted were assumed to be independent, and were assumed to be positive with probability p on any subsequent episode.

2.4.2 Isolation effectiveness

In addition to estimating transmission parameters $a_0, \ldots, a_m$, we derived an estimate for the effectiveness of isolation in combination with decolonisation treatment. Using model 2 described previously, there are various different functions which may be used to measure the effectiveness of isolation:

1. The relative risk

$$E_{\text{iso}} = \frac{P(\text{colonisation, given 1 col. patient in isolation})}{P(\text{colonisation, given 1 col. patient not in isolation})} = \frac{1 - e^{-a_0 - a_2}}{1 - e^{-a_0 - a_1}}.$$

   A beneficial effect from isolation is indicated by $E_{\text{iso}} < 1$, with the approximate reduction given by $100 \times (1 - E_{\text{iso}})\%$.

2. Similarly, we can consider the ratio $B = a_2/a_1$, which summarises the difference in transmission potential between an unisolated colonised patient and one in isolation. This is approximately similar to the relative risk when the background transmission rate is low.

3. We can compare the probabilities of avoiding colonisation in each setting:

$$\begin{aligned} A &= \frac{P(\text{avoid colonisation, given 1 col. patient in isolation})}{P(\text{avoid colonisation, given 1 col. patient not in isolation})} \\ &= \frac{e^{-a_0 - a_2}}{e^{-a_0 - a_1}} = e^{a_1 - a_2}. \end{aligned}$$

   This represents the change in susceptibility as a result of isolation. A beneficial effect is demonstrated with $A > 1$.

4. Since a_0, a_1 and a_2 typically take small values, an approximation of A is given by $D = a_1 - a_2$, the difference between the transmission rates.

5. The posterior probability $P(a_1 < a_2)$ can easily be assessed from MCMC output, and can indicate isolation effectiveness, but gives no indication about the magnitude of any potential reduction.

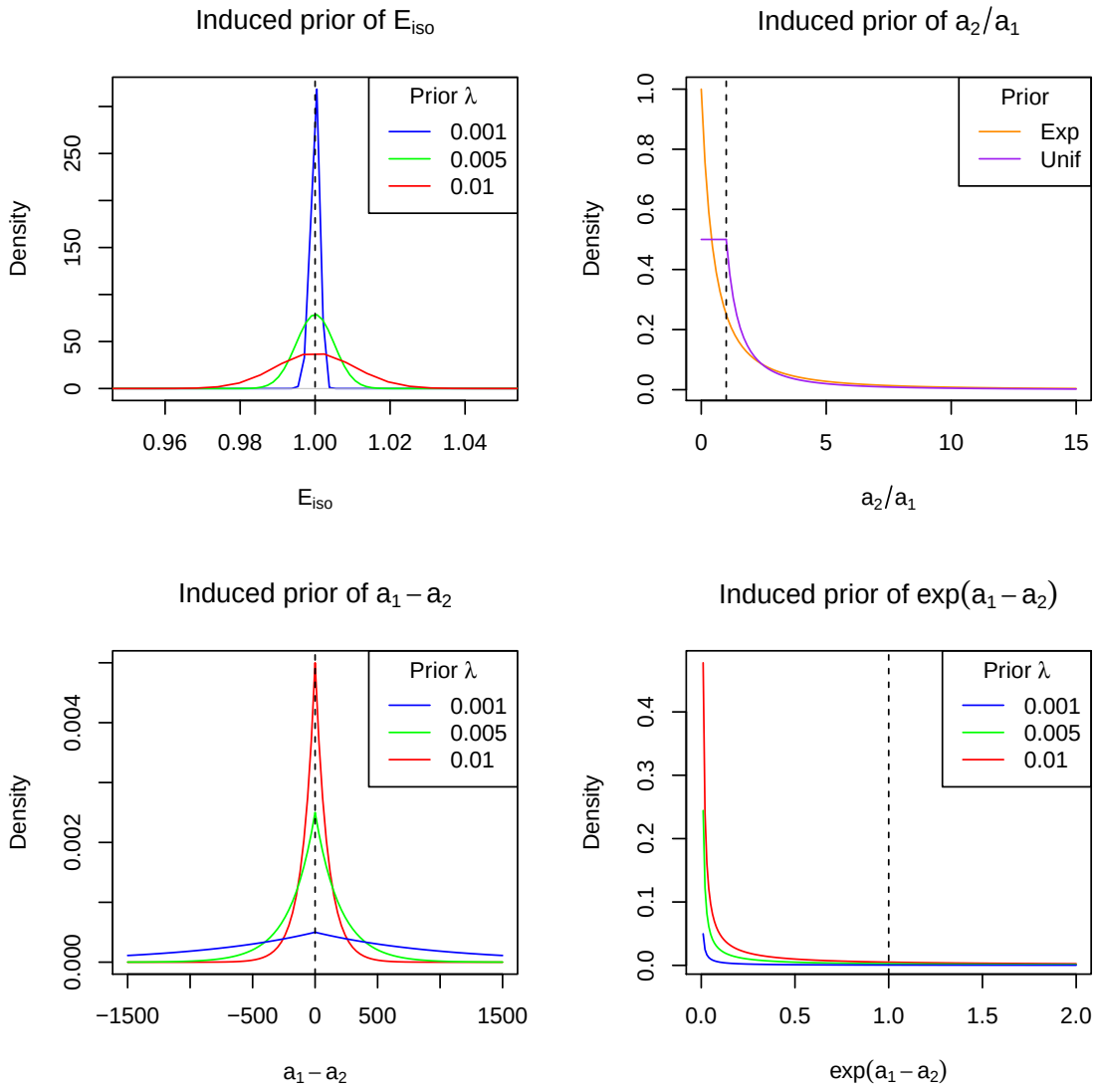

Figure 2.1: Induced priors for various isolation effectiveness functions, where transmission parameters are exponentially distributed with rate λ *a priori*, unless otherwise stated. The black dashed line on each plot indicates the 'no effect' value of the statistic. Top left is the prior distribution of E_{iso}, which becomes more informative as transmission priors become more diffuse. Top right is the prior for B, where transmission parameters are exponentially or uniformly distributed. This distribution is unaffected by the specification of the Exponential (or uniform) transmission priors. Bottom left is D, which has a Laplace distribution with exponentially distributed transmission parameters. Bottom right is A, which has a log-Laplace distribution *a priori*.

The study on isolation effectiveness by Kypraios et al. reported the ratio of transmission parameters, B, as well as the probability $P(a_1 < a_2)$ [52], while Forrester et al. reported the difference in parameter values, D [59].

A function of parameters $f(\theta)$ will have an induced prior distribution, dependent on the prior distributions assigned to θ. While the prior distributions for θ may be uninformative, the induced prior for $f(\theta)$, a function of these parameters, does not necessarily inherit this property. Indeed, for some functions, the converse is true — the more uninformative the priors for θ, the greater the *a priori* certainty of $f(\theta)$. Figure 2.1 shows the induced prior distributions of four of the isolation effectiveness functions described above, for various prior assumptions for the transmission parameters. Two of these functions are symmetrically distributed around the value indicating no isolation effect — E_{iso} and $D = a_1 - a_2$. The other two exhibit a prior mode for either positive or negative isolation effectiveness. The function $B = a_2/a_1$ is not affected by the informativeness of a_1 and a_2; in figure 2.1 we show its prior distribution under either exponentially or uniformly distributed transmission parameters. D appears to be the least informative measure of isolation effectiveness, although it is of little use in quantifying the effect.

While each of these measures provide some information on the presence of any effect, we used the relative risk, E_{iso}, to report the percentage of transmission reduction associated with isolation precautions. The induced prior of this function is informative, so we additionally conducted a sensitivity analysis to determine the effect of λ on E_{iso}.

Since decolonisation treatment was almost always initiated at the same time as isolation in this study, it was not possible to consider the effects of these two interventions separately, so we calculated the combined effect.

2.4.3 Assumptions

A number of assumptions were made when running this analysis:

1. In constructing a discrete time model, it is assumed that colonised patients contributed to the colonised population on the day after colonisation, or for importations, from the day of admission. For highly transmissible pathogens, this assumption may be inappropriate, as we eliminate the possibility of onward transmission in the initial period of colonisation. However, for bacterial pathogens such as MRSA, onward transmission is less likely in the very early stages of

colonisation, as the bacterial population is in a 'lag phase', during which the cells adapt to a new environment, and little growth occurs [168].

2. For the purposes of calculating the daily population count, the admission and isolation entry times are assumed to have occurred at the start of each day. Discharge and isolation exit times are assumed to have occurred at the end of the day. Discrete-time analysis means that colonisation events occurring on a particular day are assumed to be independent. A negative result on the day of colonisation is considered to be a false negative result.

3. Patients who were colonised with MRSA remained so for the remainder of their stay on a study ward. Carriage time of MRSA is typically long; the median length of carriage has been estimated at 8.5 months [169], while Robicsek et al. found that 48% of patients colonised with MRSA were still colonised after a year, and 21% after four years [170]. Table 2.2 shows that the typical length of stay is short, relative to such carriage times. This assumption is violated if carriage is cleared during the colonised patient's episode, or the transmission potential is reduced through the use of decolonisation therapy or antibiotics. This may result in the overestimation of colonisation pressure at any given time. No distinction was made between patients with an MRSA infection, and asymptomatically colonised patients, in terms of potential to transmit to susceptible individuals.

4. We did not explicitly model contact patterns between patients and HCWs, or direct patient-to-patient interactions, assuming that all susceptible patients were exposed to the same colonization pressure on a given day, and faced the same risk of acquisition. Similarly, we assumed that compliance with barrier precautions and the application of decolonisation treatment were the same for all patients within a ward, regardless of isolation type.

5. Colonisation was judged to be the presence of bacteria at the screening sites used in the clinical trial; nose, axilla, or groin. In this analysis, we did not account for colonisation at other sites. In section 2.7, we investigate additional observations, taken at various body sites from high-risk individuals suspected of carriage.

6. Test specificity was assumed to be 100%, meaning that false positive results were assumed not to be possible. Incorporating both sensitivity and sensitivity parameters in a model may cause identifiability issues. Experimental results indicate the specificity of screening tests to be close to 100% [138].

2.4.4 Likelihood function

Suppose a total of n patient admissions to a hospital ward are observed over some study period, labelled $\{1, \ldots, n\}$. A patient j enters the hospital at time t_j^a, and is discharged at time t_j^d. An individual j receives a set of ν_j (positive or negative) screening results $X_j = X_{j,1}, \ldots, X_{j,\nu_j}$, taken at screening times $t^x = \{t_{j,1}^x, \ldots, t_{j,\nu_j}^x\}$. If $\nu_j = 0$, patient j has no swab results, and there is no information on the disease status of this individual.

Each patient is, independently of all other patients, admitted to the hospital in a colonised state with probability p. Since MRSA carriage is asymptomatic, we rely on patient screening tests to provide an insight as to an individual's state. Screening tests detect MRSA carriage in colonised patients with probability z (test sensitivity); that is, the probability of a false negative result is $1 - z$. It is assumed that there is no chance of false positive results, that is, test specificity is 100%. We define the vector of parameters associate with model m to be $\theta_m = \{p, z, a_0, a_m\}$.

Let $t^c = \{t_1^c, \ldots, t_n^c\}$ be the set of unobserved colonisation times for patients, where t_j^c, takes a value between t_j^a and t_j^d for a patient j who is ever colonised, and $t_j^c = \infty$ if the patient remains susceptible throughout their stay. Let $\phi = \{\phi_1, \ldots, \phi_n\}$ be markers for importation: if a patient j is positive on admission, we set $\phi_j = 1$ and $t_j^c = t_j^a$, otherwise $\phi_j = 0$. Suppose that the status of all patients is known at all times, so that t^c and ϕ are also known. Let Z be the observed data which are not directly involved in the stochastic model — that is, the admission and discharge times, and isolation entry and exit times, as well as the status of patients at time $t = 0$. In our analysis, we assumed that the initial population on day 1 is zero. The joint likelihood of the swab data (X) and transmission dynamics (t^c, ϕ) is

$$\pi(X, t^c, \phi | \theta_m, Z) = \pi(X | t^c, \phi, \theta_m, Z) \pi(\phi | \theta_m, Z) \pi(t^c | \phi, \theta_m, Z). \tag{2.4.1}$$

For convenience, Z is omitted from subsequent notation, but continue to condition on these data. The first product term in the joint likelihood (2.4.1) describes the imperfect observation of the transmission dynamics, given as

$$\pi(X | t^c, \phi, \theta_m) = z^{TP(X)} (1 - z)^{FN(X, t^c)}, \tag{2.4.2}$$

where $TP(X)$ is the total number of true positive swab results, and $FN(X, t^c)$ is the total number of false negative results given the colonisation times t^c. Under the assumption of no loss of carriage, any negative result occurring after the time of colonisation is considered a false negative. We assumed that false positives are not possible, so $TP(X)$ is

directly observable from the data, and is not dependent on t^c. The second component in equation (2.4.1) describes the probability of the set of importations, given importation probability p:

$$\pi(\phi|\theta) = p^{\sum_i \phi_i}(1-p)^{n-\sum_i \phi_i}. \tag{2.4.3}$$

Finally, the transmission model is represented in the last term in the joint likelihood (2.4.1);

$$\pi(t^c|\phi,\theta) = \prod_{i=1}^{n}\Big[\mathbf{1}_{t_i^c=t_i^a} + \mathbf{1}_{t_i^c\neq t_i^a}\exp\big(-\sum_{t=t_i^a}^{\min(t_i^c-1,t_i^d)} q_m(t)\big)\Big]\prod_{\substack{j:t_j^c\neq\infty\\ \phi_j=0}}(1-e^{-q_m(t_j^c)}), \tag{2.4.4}$$

where t_i^c is constrained to take a value in $\{t_i^a,\ldots,t_i^d\}\cup\{\infty\}$, and transmission parameters $a_0,\ldots,a_m$ are constrained above zero. This represents the probability of avoiding colonisation for each susceptible patient during their stay (while $t < t_i^c$), as well as the probability of transmission for those individuals who acquire MRSA on the ward. Equation (2.4.4) can be considered a product of contributions from each patient. The contribution of patient j depends on to which of the following four cases they belong:

1. The patient remains susceptible throughout. As such, their contribution is

$$\exp\big(-\sum_{t=t_j^a}^{t_j^d} q_m(t)\big).$$

2. The patient is admitted to the ward positive. In this case, the patient makes no contribution to equation (2.4.4).

3. The patient acquires MRSA on their first day. Then the contribution is

$$1-e^{-q_m(t_j^a)}.$$

4. The patient becomes colonised after their first day. Then the contribution is

$$\exp\big(-\sum_{t=t_j^a}^{t_j^c-1} q_m(t)\big)(1-e^{-q_m(t_j^c)}).$$

In the fully-observed scenario, this likelihood may be evaluated, however, transmission dynamics are typically unobserved. Summing over all possible colonisation times results in intractability in all but the simplest cases. In order to overcome this, the parameter space was augmented with the unobserved colonisation times, and sampled across this space using an MCMC algorithm.

2.4.5 Bayesian framework

Patient screening data were analysed in a Bayesian framework, using a data-augmented MCMC algorithm. The parameter space was augmented with the unobserved data $A = \{t^c, \phi\}$, comprising the set of colonisation times and admission statuses. It is of interest to explore the posterior density $\pi(A, \theta|X)$, which, by Bayes' Theorem, is

$$\pi(A, \theta|X) \propto \pi(X|A, \theta)\pi(A, \theta) = \pi(X|A, \theta)\pi(A|\theta)\pi(\theta),$$

where $\pi(X|A, \theta)\pi(A|\theta)$ is the likelihood of the observed and augmented data given the parameters θ, and $\pi(\theta)$ is the joint prior distribution of the parameters. With known values for colonisation times, the likelihood becomes tractable, and so by treating the augmented data as parameters to be estimated, we may explore the posterior distribution.

At each iteration, the data-augmented MCMC algorithm samples values of θ, as well as the augmented data set A. This procedure is similar to previous data-augmentation approaches [52, 94, 126], although we present here a discrete-time version. The data augmentation procedure takes into account our uncertainty as to both the time and number of colonisation events, by proposing to add, delete or move elements from the set A. This design allows us to account for the unobserved patient colonisation times, crucial to the calculation of the population of colonised patients at any given time.

2.4.5.1 Prior distributions

We aimed to sample from the posterior distribution of the parameters p (probability of carriage on admission), z (culture swab sensitivity), and the transmission parameters for model m, $a_0, \ldots, a_m$. Prior distributions for these parameters were set as follows:

$$p, z \sim \text{Beta}(\alpha, \beta),$$
$$a_0, \ldots, a_m \sim \text{Exp}(\lambda),$$

where $\text{Exp}(\lambda)$ is the exponential distribution, with probability density function

$$f(x; \lambda) = \lambda e^{-\lambda x},$$

and $\text{Beta}(\alpha, \beta)$ is the beta distribution with probability density function

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

where $B(\alpha, \beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$ is the beta function. The prior distributions of our parameters were independent, and uninformative where possible; we set $\alpha = \beta = 1$ for the beta prior distributions, and $\lambda = 10^{-3}$ for the exponential priors. In order to ensure results were robust to the choice of priors, a sensitivity analysis was also conducted, results of which are presented later, in section 2.5.1.3.

2.4.5.2 Sampling model parameters

We may derive the full conditional distribution for the parameter p as follows:

$$\begin{aligned}\pi(p|\theta_{-p}, A, X) &\propto \pi(\theta_{-p}, A, X|p)\pi(p) \\ &\propto \pi(X|A, \theta)\pi(A|\theta)\pi(\theta) \\ &\propto \pi(X|A, \theta)\pi(p) \\ &\propto p^{\sum_i \phi_i}(1-p)^{(n-\sum_i \phi_i)}\pi(p),\end{aligned}$$

where θ_{-k} indicates the vector θ without the element k. From this, it follows that the probability of colonisation on admission, p, may be sampled directly from a beta distribution

$$p|\theta_{-p} \sim \text{Beta}(\alpha + \sum_i \phi_i, \beta + n - \sum_i \phi_i).$$

Similarly, the sensitivity, z, may be sampled from a beta distribution

$$z|\theta_{-z} \sim \text{Beta}(\alpha + TP(X), \beta + FN(X, A)).$$

Since the parameters p and z may be sampled directly from known distributions, we used a Gibbs step to do this.

The posterior distributions for the transmission parameters $a_0, \ldots, a_m$ are non-standard, and we used a Metropolis algorithm to derive samples from these. We used a Normal proposal distribution $q(\cdot|\theta^{(i)})$ with mean $\theta^{(i)}$ and variance σ. We ran a simple adaptive algorithm in which the acceptance rate is measured every 1000 iterations, and the variance σ is adjusted if necessary to maintain an acceptance rate between 0.2 and 0.5. In practice, after a few adjustments, the variances remained constant throughout the algorithm.

2.4.5.3 Sampling augmented data

Having sampled a new set of parameters θ, we update the augmented data. In this process, we update the importation marker, ϕ, and/or the time of colonisation, t^c by

choosing to add, move or delete a patient's colonisation time. Let ϕ^* and t^{c*} be the proposed new values for ϕ and t^c, and define $A^* = \{\phi^*, t^{c*}\}$. We assign a move probability ratio, q_{A,A^*} to each move,

$$q_{A,A^*} = \frac{P(A^* \to A)}{P(A \to A^*)},$$

where $P(A \to A^*)$ is the probability of proposing A^*, given the current dataset A.

Let v_s be the number of patients with no positive swabs, v_a be the number of patients for whom a colonisation time has been added by the algorithm, and v_q be the number of patients with a finite colonisation time. Note that v_s is a fixed value independent of the current value of A, whereas v_a and v_q both depend on the augmented data, and are updated each iteration of the algorithm. With equal probability, one of the following three moves is made:

- **Move a colonisation time.** Select at random one of the v_q patients who have MRSA at some point during their ward episode. With probability w, assigned pre-analysis, the patient is assumed to have been admitted positive. In this case, we set $\phi_j^* = 1$ and $t_j^{c*} = t_j^a$. Otherwise, we set a new colonisation time during the patient's ward episode at random. Let l_j be the the last point at which a patient can become colonised — the time of the first positive swab, if applicable, or the time of discharge if the patient never has a positive swab. We set $\phi_j^* = 0$ and sample a value uniformly from $\{t_j^a, \ldots, l_j\}$ for the colonisation time t_j^{c*}. The proposal probability ratio q_{A,A^*} is given as follows:
  - Change acquisition time to new acquisition time: $q_{A,A^*} = 1$.
  - Change acquisition time to importation: $q_{A,A^*} = \frac{1-w}{w(l_j - t_j^a + 1)}$.
  - Change importation to acquisition time: $q_{A,A^*} = \frac{w(l_j - t_j^a + 1)}{1-w}$.
  - Remain an importation (no change): $q_{A,A^*} = 1$.
- **Add a colonisation time.** Choose at random one of the $v_s - v_a$ patients who neither acquire nor import the pathogen. If $v_s - v_a = 0$, no move is made. With probability w an importation is added; in this case we set $\phi_j^* = 1$. Otherwise, draw a random sample from the discrete uniform distribution from admission to discharge time, and assign this to the new colonisation time t_j^{c*}.
- **Remove a colonisation time.** Choose one of the v_a patients who have previously had a colonisation time added. If $v_a = 0$, no move is made.

Having established the augmented data move mechanisms, the probability ratios q_{A,A^*} for adding or removing colonisation times may be given as follows:

	Importation	Acquisition
Add	$\frac{v_s - v_a}{w(v_a+1)}$	$\frac{(v_s-v_a)(t_j^d - t_j^a+1)}{(1-w)(v_a+1)}$
Remove	$\frac{v_a w}{v_s - v_a + 1}$	$\frac{v_a(1-w)}{(t_j^d - t_j^a+1)(v_s-v_a+1)}$

Having proposed a new augmented dataset A^* and calculated q_{A,A^*}, we accept the proposed move with probability

$$\min\left(1, \frac{\pi(X|A^*,\theta)\pi(A^*|\theta)}{\pi(X|A,\theta)\pi(A|\theta)} q_{A,A^*}\right). \tag{2.4.5}$$

The data augmentation process here is a form of the reversible jump MCMC algorithm [124]. As described in section 1.5.2.7, the augmented data transformation function $P(A \to A^*)$ generates proposals A^* independent to the current value A. This means that the Jacobian determinant present in the RJMCMC acceptance rate is equal to 1, and may be ignored.

The MCMC algorithm was run for 120,000 iterations, with a burn-in of 20,000, on the data collected from each of the ten study wards. The augmented data sampling step of the algorithm was repeated ten times per iteration to improve the speed of mixing. The value of w was set to be 0.3 in the data augmentation step. The algorithm was written in `C`, and the analysis of the output was performed in `R 2.10.1` [171].

2.4.5.4 Pooled estimates

In order to obtain overall parameter estimates across the ten hospital general wards, a random effects meta-analysis was used to pool the individual ward estimates.

Suppose estimates $x_1, \ldots, x_n$ of a parameter μ are derived from n different sources (studies, datasets, etc.). A random effects meta-analysis assumes that each estimate can be expressed as $x_i = \alpha_i + \epsilon_i$, where α_i is the true value for study i, and ϵ_i is the associated error, with variance σ_i. Furthermore, the true value for any given study or dataset i, α_i, is assumed to vary from the overall true value μ, due to random effects on the study as a whole. This means that α_i can be expressed as $\alpha_i = \mu + \delta_i$, where δ_i is the between-study error, with variance τ. As a result, the estimates $x_1, \ldots, x_n$ can be considered as samples from a distribution with mean μ, and variance $\tau + \sigma_i$. With known within-study and between-study variances $\sigma_1, \ldots, \sigma_n$ and τ respectively, a pooled esti-

mate for μ incorporating the n sources, can be given as

$$\hat{\mu} = \frac{\sum_{i=1}^{n} W_i x_i}{\sum_{i=1}^{n} W_i},$$

where $W_i = \frac{1}{\sigma_i + \tau}$, in which each estimate x_i is weighted by W_i, the reciprocal of the combined within and between study variance. Typically, these variances are not observed, and as such, estimates must be used instead. DerSimonian and Laird described methods to calculate a pooled estimate $\hat{\mu}$ with estimated within-study and between-study variances [172].

In this analysis, a random-effects meta-analysis was implemented to derive pooled parameter estimates using the `rmeta` package [173] in `R` [171], which utilises DerSimonian and Laird's method to calculate weights.

2.4.6 Goodness of fit

Model fit was assessed by analysing the posterior predictive distribution of specific admission-discharge swab pairs (positive-positive, positive-negative, negative-positive, negative-negative). A total of 5000 datasets $\tilde{X}$ were simulated, based on parameters drawn from the posterior densities derived from the data-augmented MCMC algorithm, counting the number of swab pairs generated each time. Let n_{xy} be the observed number of (x, y) swab pairs, and $\widetilde{n_{xy}}$ be the number of simulated swab pairs. For each of the four possible swab pairs, we calculated the quantile $q_{xy} = P(\widetilde{n_{xy}} < n_{xy})$, sometimes referred to as the posterior predictive p-value [130]. Any extreme values of q_{xy} (close to 0 or 1) are supportive of a lack of fit, indicating that the model would be unlikely to predict the data we have observed. This analysis was repeated for each of the ten study wards.

2.4.7 Pseudolikelihood approximation

While the Bayesian approach offers great benefits in terms of flexibility, the data-augmented MCMC algorithm can be time-consuming and difficult to implement. It is certainly of interest to know whether simpler methods can be employed to attain reasonable parameter estimates at considerably less computational expense.

Observing the transmission dynamics accurately is not possible, especially for asymptomatic colonisations caused by pathogens such as MRSA, which results in an intractable likelihood, requiring integration over all possible colonisation times. How-

ever, by assuming patient episodes are independent, and that the colonised population is fixed, the true likelihood may be approximated. The pseudolikelihood of data $X = \{X_1, \ldots, X_n\}$, given parameters θ is

$$\pi^*(X|\theta) = \prod_{i=1}^{n} \pi(X_i|X_{-i}, \theta),$$

where $\pi(X_i|X_{-i}, \theta)$ is the likelihood of observation X_i, given the remainder of the observations. The pseudolikelihood matches the likelihood function in the event that $X_1, \ldots, X_n$ are independent. The $\tilde{\theta}$ that maximises π^* is then the maximum pseudolikelihood estimate (MPE).

Hospital patient data typically comprises a set of patient episodes much shorter than the overall data collection period, and therefore any given episode will coincide with only a small subset of the other observations. Only overlapping patient episodes are dependent, and a low degree of dependence may indicate that the pseudolikelihood is a reasonable approximation.

2.4.7.1 Model & assumptions

The disease state of a patient depends on other individuals via the transmission rate, $q(t)$. In order to calculate the pseudolikelihood, each patient must be assumed to be independent, and as such, we must assume that we know $q(t)$ for all t. This can be achieved by making an assumption about colonisation times, based on observed swabs. For example:

1. Set the colonisation time to be the midpoint between the first positive result and the previous negative result. If no such negative result exists, we adopt the admission time as the lower bound of this interval.

2. We can get a lower bound on the observed colonised population by assuming colonisations occurred just before the first positive swab.

3. Similarly, we can get an observed upper bound by assuming any colonisation observed was present on admission.

These assumptions — made only in order to fix colonisation population at all times, and not used in modelling transmission — allow MPEs to be calculated. The more frequent (and more accurate) the screening, the closer we approach to the true number of colonised individuals at any given time. We used the first of the above assumptions

to calculate $q(t)$, but additionally ran the analysis for the alternative proposals, in order to determine the sensitivity of the estimates to this assumption.

Let f_j be the time of patient j's first positive screen. If this patient has no positive results, set $f_j = \infty$. Using the same notation as previously, we suppose that the likelihood contribution of patient i, given all other patients, can be written as

$$\begin{aligned}
\pi(X_i|X_{-i},\theta) =& P(\text{observed data, colonised on admission}) \\
&+ \sum_{t=t_j^a}^{t_j^d} P(\text{observed data, colonised on day t}) \\
&+ P(\text{observed data, never colonised}) \\
=& pz^{TP_i}(1-z)^{FN_i(t_j^a)} \\
&+ \mathbf{1}_{f_i \neq \infty}(1-p)\Big[(1-e^{-q_m(t_i^a)})z^{TP_i}(1-z)^{FN_i(t_i^a)} + \\
&\sum_{t=t_j^a}^{\min(f_i,t_i^d)} \exp(-\sum_{i=t_i^a}^{t-1} q_m(i))(1-e^{-q_m(t)})z^{TP_i}(1-z)^{FN_i(t)}\Big] \\
&+ \mathbf{1}_{f_i=\infty}(1-p)\exp(-\sum_{i=t_i^a}^{t_j^d} q_m(i)),
\end{aligned}$$

where TP_i and $FN_i(t)$ are the number of true positive and false negative screening results for patient i, given colonisation occurred on day t. The pseudolikelihood is then equal to the product of the contribution from each patient.

2.4.7.2 Uncertainty

A Nelder-Mead optimisation algorithm [100, 101] was used, implemented using the `nlm` function in `R` [171], to derive MPEs. Confidence intervals can be derived from the Hessian matrix returned by this function. The inverse of the Hessian matrix is asymptotically equal to the covariance matrix for MLEs, and may therefore be used to estimate confidence intervals. This does not necessarily hold when optimising a pseudolikelihood function, and may potentially provide a poor approximation of the confidence interval.

Estimates of variance derived in this method should be treated with caution, as inaccurate values may be returned if the parameter estimate is close to the boundary of the support. This may be the case in our model, where transmission parameters a_0, a_1, a_2 are constrained above zero, but take very low values. In order to avoid this issue, we transformed the parameters and optimised the pseudolikelihood with respect to the

log of the transmission parameters, in order to allow unconstrained numerical optimisation.

An alternative method to generate confidence intervals is via parametric bootstrap [110], described earlier in section 1.5.1.3. The maximum pseudolikelihood estimate $\hat{\theta}$ is used to generate n datasets $\tilde{X}_1, \ldots, \tilde{X}_n$, each of which is then used to derive n further estimates $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$. A 95% bootstrap confidence interval may then be approximated by taking the 2.5% and 97.5% quantiles of $\{\tilde{\theta}_1, \ldots, \tilde{\theta}_n\}$.

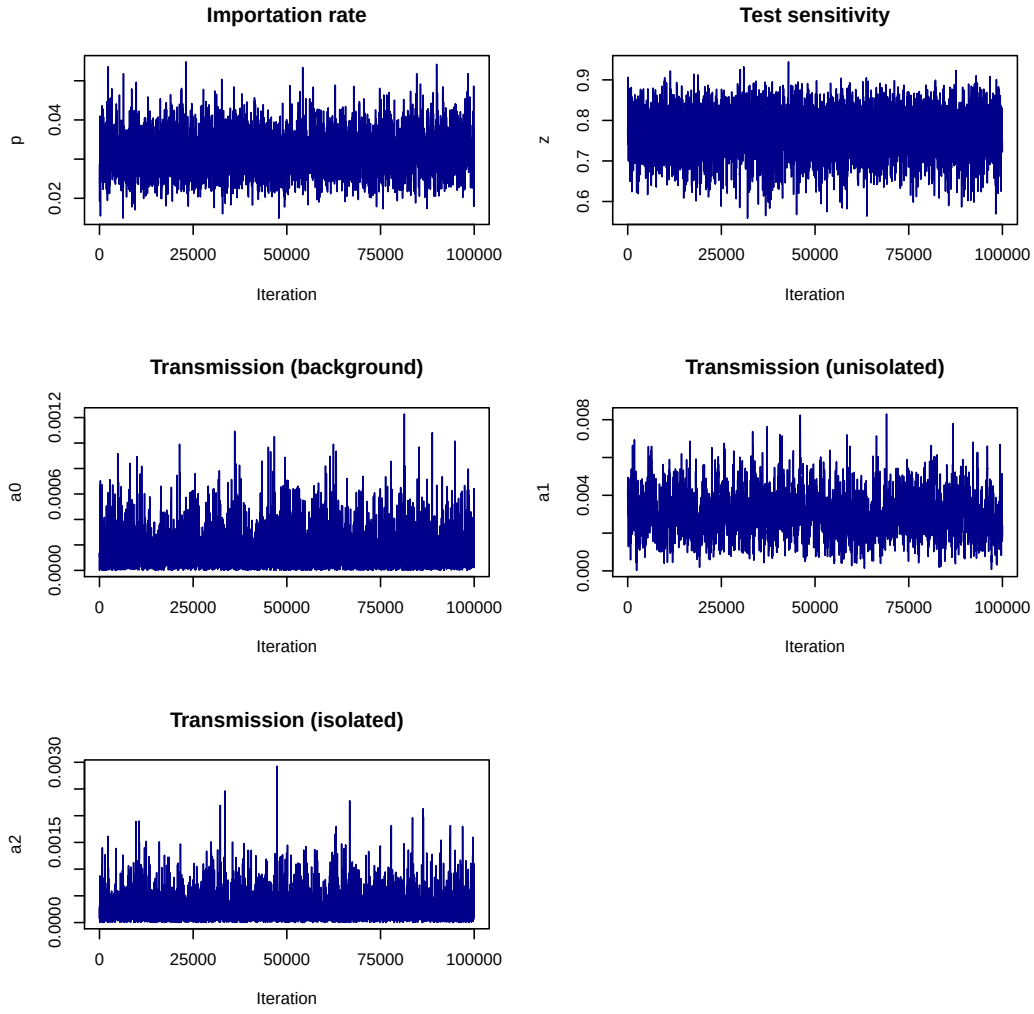

Figure 2.2: MCMC samples for the parameters θ using data from ward 1, under model 2. The burn-in period has been disregarded, and the samples thinned by a factor of 10.

2.5 Results

We now present results from the data-augmented MCMC analysis of MRSA transmission in the ten hospital general wards, followed by pseudolikelihood estimates. While we present results for each of the three models described earlier, our primary interest lies in model 2. It was found that estimates derived from the more complex model (model 3, in which isolation is broken down by type) were associated with a great deal of uncertainty, due to the sparsity of the data involved. Unless otherwise stated, results presented here have been derived using model 2. Estimates for the non-transmission parameters (p and z) remained similar between models.

2.5.1 MCMC approach

2.5.1.1 Model parameters

Convergence and mixing were monitored for each implementation of the MCMC algorithm via visual inspection and comparison of chains with various starting points. Trace plots for ward 1 are shown in figure 2.2.

Prevalence on admission, p, varied considerably between the wards (figure 2.3), as might be expected due to the different patient types admitted to each ward. Estimates ranged from 3% to 16%, with the highest importation rates estimated for the two elderly care wards (2 and 6). These wards also saw the highest rate of admission from elsewhere in the hospital, and many of the patients positive on admission may have previously been recorded as positive in other wards. Estimates for sensitivity, z, ranged from 58% to 86% on individual wards (shown in figure 2.4). The pooled ward estimate, derived from a random-effects meta-analysis, was 77% (95% CI: 72%, 82%). In addition, we estimated the proportion of colonised days spent out of isolation. This value ranged from 40% to 65% between the wards, and we estimated the pooled value to be 54% (47%, 60%).

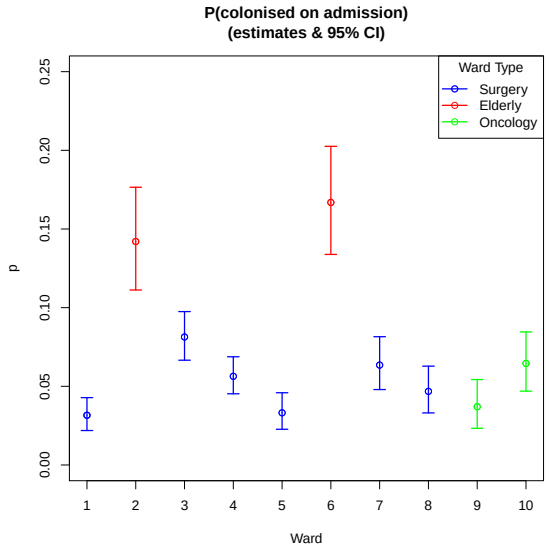

Figure 2.3: Ward estimates for the parameter p, probability of colonisation at admission, together with 95% equitailed credible intervals calculated from MCMC samples.

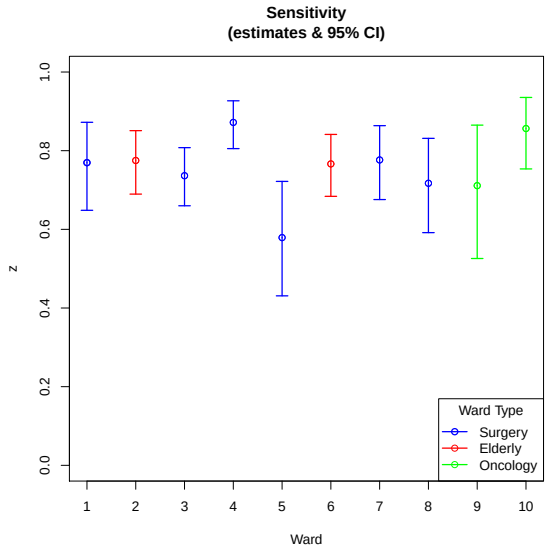

Figure 2.4: Ward estimates for the parameter z, test sensitivity, together with 95% equitailed credible intervals calculated from MCMC samples.

Pooled estimates for the transmission parameters for each of the models are shown in table 2.3. Background transmission was found to have a relatively minor effect on the MRSA acquisition rate in this setting. The pooled estimate of a_0 indicates that an expected 0.23 colonisations per 1000 patient days occur due to a background effect. This is a small fraction of the colonisations expected to arise due to transmission from other patients (table 2.4). Transmission rates varied between the study wards, ranging between one and five expected acquisition events per 1000 patient days (table 2.4). High rates were estimated for the elderly care wards, while oncology wards had the lowest risk of transmission. Expected transmission rates were lowest according to model 1, and highest in model 3, but were similar overall.

The posterior distributions for each parameter in model 2 are shown in figure 2.5, along with scatter plots and correlation coefficients of MCMC samples between each pair of parameters. Two wards are shown here; a high prevalence elderly care ward (ward 2) and a surgical ward with average prevalence (ward 3). This illustrates the skewness of the transmission parameter posterior distributions, but shows little correlation between parameters. However, as expected, lower estimates for sensitivity, z, are associated with higher estimates for the prevalence of carriage on admission, p. Furthermore, we found that correlation in the log scale to not exceed ± 0.3.

Pooled transmission parameter estimates for each model

Model	Transmission rate, $q(t)$	a_0	a_1	a_2	a_3
1	$a_0 + a_1 C(t)$	0.000275	0.00066	—	—
2	$a_0 + a_1 C_N(t) + a_2 C_I(t)$	0.000230	0.00125	0.000275	—
3	$a_0 + a_1 C_N(t) + a_2 C_W(t) + a_3 C_{SR}(t)$	0.000229	0.00106	0.000737	0.000321

Table 2.3: Transmission parameter estimates derived from a meta-analysis of individual ward posterior median estimates, according to each model, where $C(t)$ is the number of colonised patients in total at time t, $C_I(t)$ of whom are isolated, and $C_N(t)$ are not. Of the isolated patients, $C_{SR}(t)$ are in a side room, and $C_W(t)$ are isolated on the open ward.

2.5.1.2 Isolation effect

Estimates for the effectiveness of isolation and decolonisation for each ward, derived from model 2, are shown in table 2.5. All but two wards show a reduction in transmission associated with isolation measures (ward 7, surgery and ward 9, oncology), and the posterior probability that isolation is effective is over 90% in six wards. A

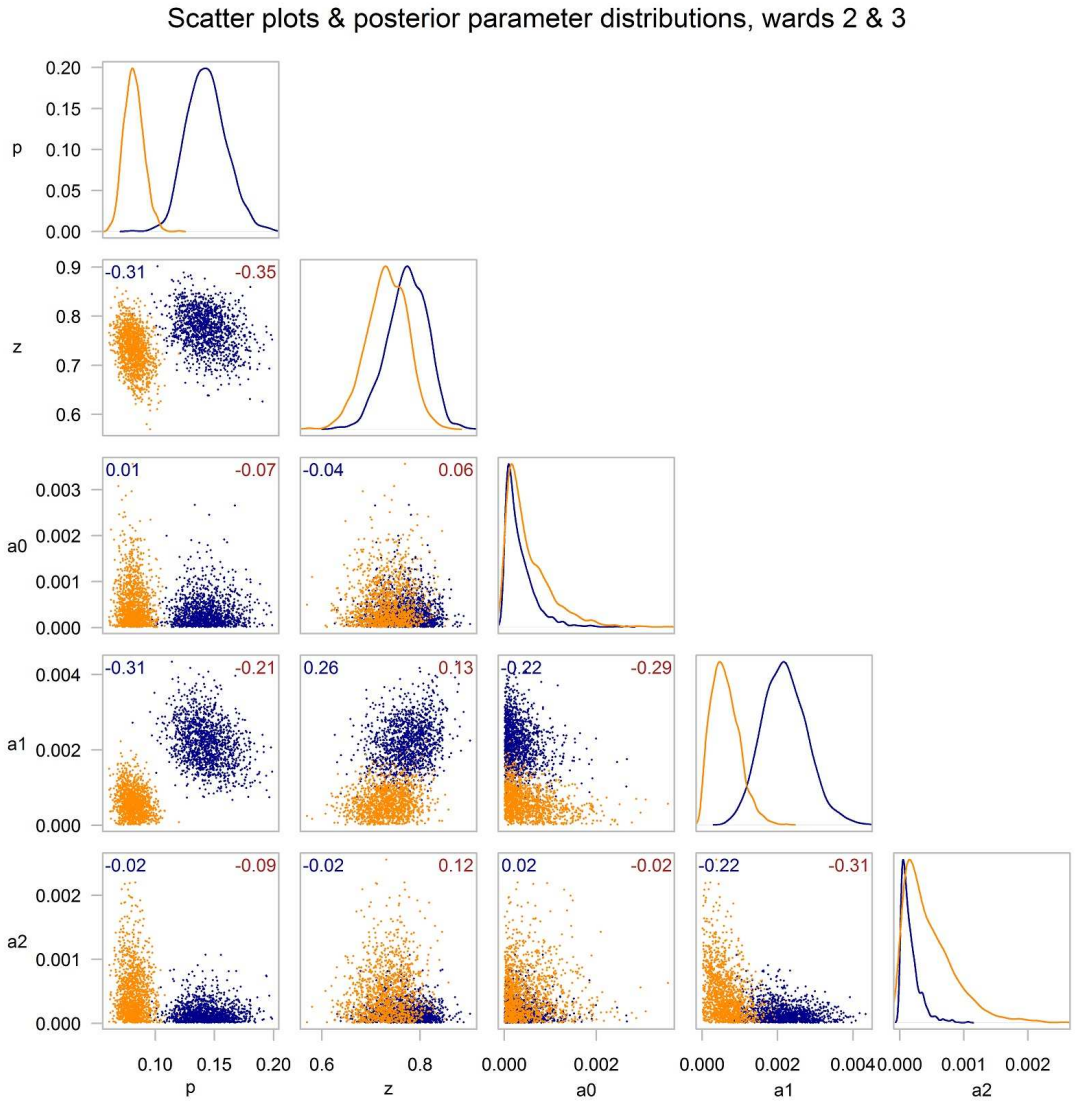

Figure 2.5: Posterior densities and pairwise MCMC sample plots for two wards, under model 2. Samples from the data-augmented MCMC algorithm are plotted for each pair of parameters, along with correlation coefficients. Posterior distributions are shown on the diagonal for each parameter. We use ward 2 (high prevalence, elderly care ward, plotted in blue) and ward 3 (average prevalence, surgical ward, plotted in orange) as examples.

meta-analysis was conducted on the log of the relative risk function, as this provided a better approximation to normality, which is assumed in the meta-analysis process. A forest plot for the isolation relative risk is shown in figure 2.6, derived from model 2. The pooled estimate for the relative risk is 0.36 (0.21, 0.63), indicating that isolation and decolonisation treatment are associated with a reduction in transmission of

		Estimated colonisation rate
Ward	Type	Expected acquisitions per 1000 patient days (95% CrI)
1	Surgery (plastics)	1.7 (0.7, 3.0)
2	Elderly care	4.7 (2.5, 7.0)
3	Surgery (urology)	2.4 (0.8, 4.5)
4	Surgery (ear, nose & throat)	3.5 (2.1, 5.1)
5	Surgery (cardiothoracic)	2.0 (0.9, 3.6)
6	Elderly care	3.8 (2.0, 5.8)
7	Surgery (vascular)	2.3 (1.3, 3.6)
8	Surgery (gastrointestinal)	2.7 (1.6, 4.2)
9	Oncology	1.3 (0.5, 2.5)
10	Oncology	1.5 (0.6, 2.6)

Table 2.4: Posterior mean estimated rate of acquisition in each ward, calculated under model 1.

approximately 64%.

The expected proportions of transmission events due to three sources (background, isolated MRSA positive patients within the same ward, unisolated MRSA positive patients within the same ward) and rate of transmission for each ward are shown in figure 2.7. We estimate that, under model 2, approximately 75% (67%, 86%) of transmission in this setting is attributable to MRSA positive patients under no isolation precautions. Unisolated colonised patients are the source of the majority of ward transmission in all but two wards (wards 7 and 9, for which we did not estimate a reduction in transmission). Background transmission was estimated to account for 9% (3%, 14%) of within-ward transmission.

Intervention effectiveness estimates were derived for each isolation type from model 3, and a pooled estimate was again calculated by performing a meta-analysis on the log-effect for each ward. The side room isolation relative risk, E_{SR}, was estimated to be 0.41 (0.23, 0.72), and the estimate of E_W, the effect of cohorting and contact precautions, was 1.47 (0.69, 3.13). Therefore, while we estimate side room isolation to be associated with a reduction in transmission (of approximately 59%), the data do not suggest that open ward isolation has a similar effect; however the uncertainty surrounding this estimate is large. Open ward isolation was used infrequently (see table 2.6) compared to side room isolation on most wards. Usage was highest in wards 2 and 6 (elderly care), and

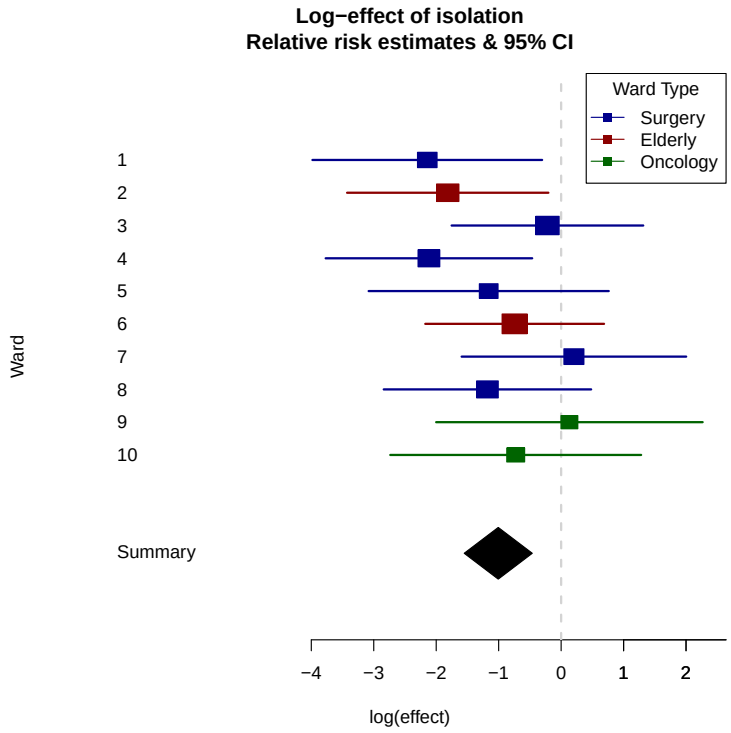

Figure 2.6: Isolation effectiveness. Estimates for the reduction in transmission rate due to any kind of isolation, together with decolonisation treatment, compared to an unisolated MRSA positive patient. These estimates were derived using model 2.

open ward isolation was estimated to have a beneficial effect in this setting; the pooled effect for open ward isolation effectiveness in elderly care wards was estimated to be 0.44 (0.15, 0.95).

2.5.1.3 Model verification

In order to assess goodness of fit for model 2, the posterior predictive distribution was estimated for the possible admission-discharge swab pairs for each of the ten wards, and the posterior predictive p-value, $q_{xy} = P(\widetilde{n_{xy}} < n_{xy})$, was derived. Figure 2.8 shows the four posterior predictive distributions for ward 3 as an example. No extreme p-values were found; that is, all p-values were contained in the interval $(0.025, 0.975)$.

A sensitivity analysis was performed to determine the impact of the transmission parameter prior rate λ on the estimate of isolation effectiveness. The prior rate was varied from $\lambda = 10^{-1}$ to $\lambda = 10^{-6}$. As mentioned earlier, this affects the induced prior of E_{iso}, and could potentially impact our results. However, it was found that found that the

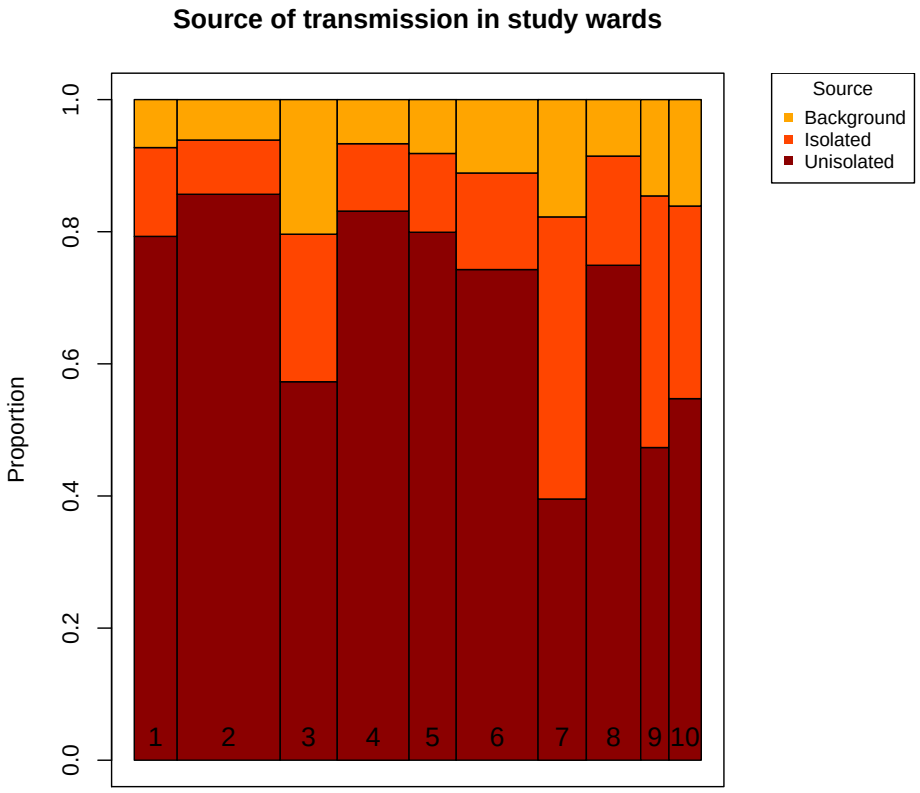

Figure 2.7: The estimated proportion and rate of transmission attributable to each source. Each column represents the transmission from a particular study ward. The widths of the columns are proportional to the estimated mean rate of transmission. The transmission rate is then split into the proportions expected from (top to bottom) background transmission, colonised patients in isolation, and colonised patients not receiving isolation precautions.

estimate was robust to changes in the prior distribution (see table 2.7).

We ran several simulation studies, in order to test the ability of the model to recover and identify the parameters well. Datasets of a similar size to the study wards were generated, and we tested scenarios with different transmission rates and importation probabilities. In the majority of cases, we found that parameters were estimated adequately. Transmission parameters were estimated less successfully in low transmission scenarios, due to a lack of data. This was an issue for all models, but particularly for model 3. In particular, we found that the parameter a_0 was often underestimated. Fig-

Isolation effectiveness estimates

Ward	E_{iso}	$P(a_1 > a_2)$
1	0.12	0.99
2	0.16	0.99
3	0.8	0.64
4	0.12	0.99
5	0.31	0.91
6	0.48	0.93
7	1.23	0.39
8	0.31	0.94
9	1.14	0.45
10	0.48	0.79

Table 2.5: Estimates for the effectiveness of isolation in combination with decolonisation treatment, E_{iso}, calculated from model 2, alongside the probability that isolation measures are effective, $P(a_1 > a_2)$. The reduction in transmission is given as $100(1 - E_{iso})\%$.

ure 2.5 also shows that a_0 is slightly correlated with the other transmission parameters. There may be a concern that the identifiability of a_0, together with slight correlations to other parameters, might impact our isolation effectiveness estimate. In order to verify this, the analysis of model 2 was repeated, with a_0 set to zero, such that the transmission rate was $q(t) = a_1 C_N(t) + a_2 C_I(t)$. There was little difference in the estimates of E_{iso} (pooled E_{iso} estimate 0.38, compared to 0.36 in the model including a_0).

2.5.2 Pseudolikelihood

We analysed the pseudolikelihood for model 2, assuming, for the purposes of the population count, that colonisations occurred at the midpoint of the known susceptible period. Optimisation using the Nelder-Mead algorithm in `R` took 30-120 seconds to produce point estimates and confidence intervals. This is much quicker than the MCMC procedure described earlier, which took between 2 and 6 hours to run 100,000 iterations. However, using a bootstrap approach to derive estimates of E_{iso}, or to generate bootstrap confidence intervals, took much longer.

Results for the prevalence on admission (p) and swab sensitivity (z) were found to be similar to those obtained in the Bayesian approach. In addition, the 95% confi-

Isolation usage and effectiveness estimates

Ward	Total patient days (colonised)	SR isolation days (colonised)	Open ward isolation days (colonised)
1	9975 (652)	1446 (311)	246 (80)
2	10734 (2095)	1474 (636)	897 (454)
3	7907 (1793)	1027 (522)	236 (89)
4	11020 (1358)	2123 (783)	46 (19)
5	8685 (657)	1048 (222)	19 (5)
6	10539 (2792)	1392 (672)	1157 (616)
7	10264 (1991)	1440 (721)	331 (68)
8	8959 (1429)	1278 (402)	394 (120)
9	8797 (622)	3924 (263)	115 (9)
10	7867 (919)	1563 (463)	103 (39)
All	94747 (14308)	16715 (4995)	3544 (1499)

Table 2.6: Total number of patient days for all wards during the study period, alongside the number of days spent in side room (SR) isolation, and open ward isolation. The numbers in brackets indicate the posterior mean estimate of the number of colonised patient days in each setting.

dence intervals, generated from the inverted Hessian matrix, were of a similar size to 95% credible intervals obtained earlier. Figure 2.9 shows estimates obtained under both approaches for each ward. There seemed to be no consistent pattern regarding over/underestimation with the pseudolikelihood approach.

MPEs for transmission parameters exhibited greater deviation from the Bayesian estimates. Several ward-level estimates, particularly for a_0, were extremely close to zero. Model parameter estimates are summarised in table 2.8.

The Nelder-Mead algorithm failed to converge when optimising the likelihood under model 3, potentially indicating over-parametrisation for the amount of data available.

The analysis was run under the assumption that the colonised population was already known, calculated by taking the time of colonisation to be the midpoint of the first observed positive result and the previous negative swab, or admission. We additionally ran the analysis assuming that colonisation occurred at the start, or the end, of this interval, in order to assess the impact of this assumption. We found that this affected our isolation effectiveness estimates, which were slightly lower when acquisition times

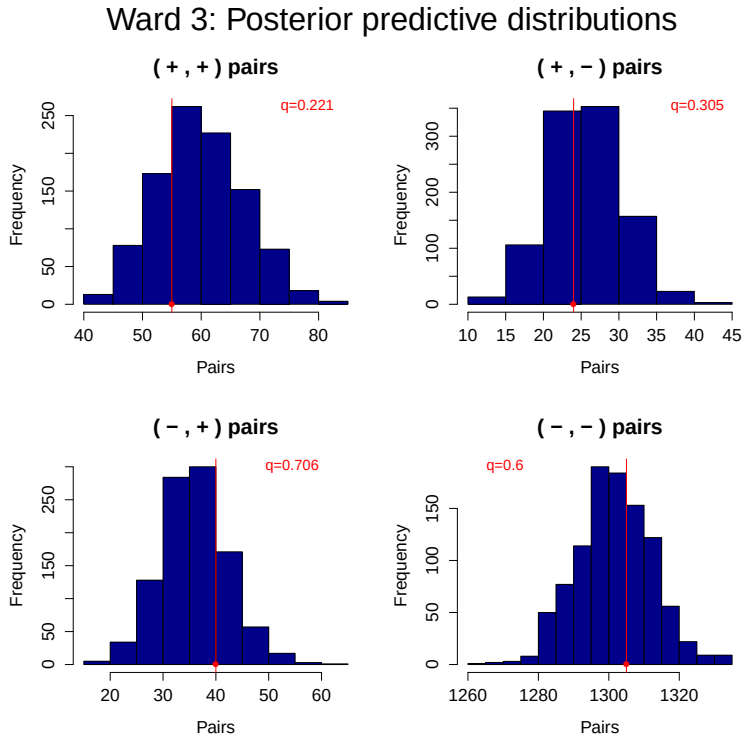

Figure 2.8: Posterior predictive distributions for given admission-discharge swab pairs, with ward 3 as an example. The red line indicates the observed number of each swab pairing, and the blue histogram represents the values obtained from 1000 simulated datasets.

were assumed to be at the start of the interval, and higher when at the end. This might be explained by the fact that isolation is likely to occur only once carriage is detected, and therefore assuming that acquisition takes place at the end of an interval is likely to underestimate the unisolated colonised population. This may in turn inflate the estimate of a_1, reducing the estimated isolation effect. The opposite effect may cause the increase in isolation effect while assuming acquisition occurs at the start of the interval.

Sensitivity of E_{iso} to transmission parameter prior

Prior rate (λ)	Pooled E_{iso}
10^{-1}	0.37 (0.21,0.65)
10^{-3}	0.36 (0.21,0.63)
10^{-6}	0.39 (0.23,0.67)

Table 2.7: The sensitivity of the estimate for isolation effectiveness to the prior assigned to the transmission parameters. Parameters are exponentially distributed *a priori* with rate λ.

Maximum pseudolikelihood parameter estimates

Ward	a_0	a_1	a_2	E_{iso}
1	1.5×10^{-8}	0.0014	8.6×10^{-6}	0.14
	(0,0.0009)	(0.0014,0.0015)	(0,0.001)	(0.03,0.64)
2	1.5×10^{-5}	0.0032	1.4×10^{-8}	0.02
	(0,0.001)	(0.0018,0.0045)	(0,0.001)	(0.01,0.06)
3	3.9×10^{-9}	0.0016	0.00021	0.25
	(0,0.0013)	(0.0016,0.0017)	(0,0.0015)	(0.02,2.73)
4	0.0012	0.0008	7.2×10^{-11}	0.60
	(0.0012,0.0013)	(0.0004,0.0013)	(0,0.0004)	(0.11,3.37)
5	4.4×10^{-9}	0.0047	0.00011	0.06
	(0,0.0017)	(0.0005,0.0089)	(0,0.0018)	(0.01,0.36)
6	5.5×10^{-10}	0.0019	0.00022	0.11
	(0,0.0011)	(0.0019,0.0020)	(0,0.0013)	(0.04,0.30)
7	1.7×10^{-8}	0.0003	0.0013	4.29
	(0,0.0007)	(0,0.0010)	(0.0012,0.0013)	(0.33,51.82)
8	2.0×10^{-9}	0.0016	0.00025	0.17
	(0,0.0011)	(0.0016,0.0017)	(0,0.0014)	(0.05,0.55)
9	3.5×10^{-8}	0.0007	0.0010	1.51
	(0,0.0010)	(0,0.0017)	(0.0010,0.0011)	(0.07,34.91)
10	2.1×10^{-10}	0.0014	0.00062	0.42
	(0,0.0011)	(0.0014,0.0015)	(0,0.0017)	(0.05,3.66)

Table 2.8: Maximum pseudolikelihood estimates for transmission parameters in model 2 for each ward, along with 95% confidence intervals derived from the inverted Hessian matrix. Estimates and 95% confidence intervals for isolation effectiveness E_{iso} were obtained from the median value of 1000 bootstrap samples.

2.6 Discussion

2.6.1 Isolation and decolonisation effectiveness

Our analysis provides strong evidence that isolation precautions in combination with decolonisation treatment are associated with a reduction in MRSA transmission in hospital general wards. There have been few studies to provide evidence for a decline in transmission associated with the use of these precautions, and indeed, little investiga-

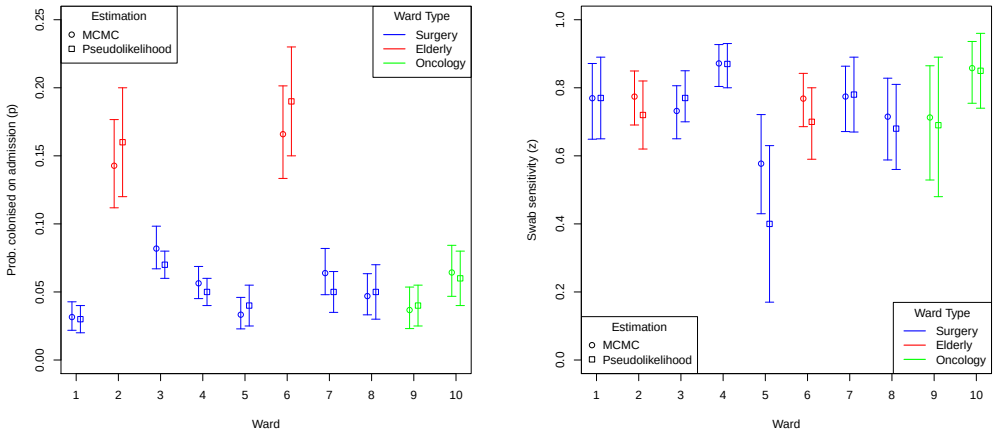

Figure 2.9: Maximum pseudolikelihood estimates alongside estimates from the MCMC procedure for prevalence on admission (left) and swab sensitivity (right). Uncertainty shown with 95% confidence intervals derived from the Hessian matrix and 95% credible intervals as appropriate.

tion into the transmission of MRSA in a general ward setting. Our analysis suggests that most transmission within these wards was due to patient-to-patient spread from unisolated MRSA positive patients.

We also investigated the separate effects of side room isolation and open ward isolation on MRSA transmission. We found a reduction of approximately 59% (28%, 77%) in transmission associated with side room isolation, but did not estimate any beneficial effect from open ward isolation. This estimate is surrounded by great uncertainty, due to the strong preference for using side room isolation whenever possible. Table 2.6 shows the number of patient days spent in each type of isolation for each ward. In most wards, just a small fraction of the total number of colonised days were spent in open ward isolation. The elderly care wards saw the highest usage of non side room isolation, as a high prevalence of MRSA in the wards regularly caused the side rooms to be full to capacity. This form of isolation was estimated to have a beneficial effect for these wards, albeit a slightly lesser effect than side room isolation. The difference in the estimates between isolation types might be explained by the fact that open ward isolation was most commonly used when the side rooms were full, which often coincided with times of high MRSA prevalence. It is possible that all forms of isolation could be less effective during times of high prevalence, due to the extra burden this places on hospital staff.

Though we found isolation and decolonisation to be associated with considerably lower transmission from MRSA carriers, our results do not tell us about the relative importance of the measures, or about the causal mechanisms responsible for the association. It is possible for example that reduction was due to increased hand hygiene amongst HCWs treating patients known to be positive. Since the majority of decolonisation treatment was initiated at the same point as isolation, the data available for individuals receiving only one intervention were sparse. A study in which interventions are staggered may provide adequate data to analyse their independent effects. Alternatively, we might consider the effect of decolonisation treatment to be non-linear. For example, decolonisation treatment may not have an immediate effect, and could be assumed to have a delayed effect in the reduction of transmissibility.

In one surgical ward and one oncology ward, isolation was estimated to have a detrimental effect on MRSA transmission, and one further surgical ward saw only a negligible beneficial effect (figure 2.6). This could be due to chance variation, as in all cases associated credible intervals are wide. However, other explanations are of course possible. For example, there are case reports of superspreading events [174] and recent modelling work has highlighted the potential for a peripatetic healthcare worker with poor hand hygiene compliance to influence the transmission dynamics greatly [175], and this may reduce the effect of isolation.

Table 2.6 shows the number of patient days spent in each form of isolation for each ward. We found over half of all colonised patient-days to be spent out of isolation. False negative swab results and delays in screening (and the processing of results) contribute to this. In most wards, just a small fraction of the total number of colonised patient-days were spent in open ward isolation. The elderly care wards saw the highest usage of non side room isolation, as a high prevalence of MRSA in these wards caused the side rooms to be full to capacity for much of the study period. Interestingly, in these two wards the effectiveness of open ward isolation was similar to that of side room isolation.

2.6.2 Modelling assumptions

Several assumptions were made in order to perform the analysis. The decision to assume positive patients remain positive during their stay was made because carriage time of MRSA is typically long; the median length of carriage has been estimated at 8.5 months [169], while Robicsek et al. found that 48% of patients colonised with MRSA

were still colonised after a year, and 21% after four years [170].

Specificity of the screening test to detect MRSA was assumed to be 100%, and independent estimates have confirmed that that specificity closely approaches this value. Two culture tests were used in the clinical trial; CHROMagar MRSA and an MRSA selective broth, which were estimated by Perry et al. to have specificities of 99.3% and 92.8% respectively, after 22-24 hours [138]. We found sensitivity estimates varied slightly across the wards. Sensitivity can depend on swabbing techniques and the time between taking the swab and testing, which may account for the differences between ward estimates. Our estimate of sensitivity is an approximation of the clinical sensitivity of the test, based on the assumption that positive patients remain positive for the duration of their stay. If individuals were actually cleared of carriage during their hospital episode, then our model will underestimate sensitivity, as we assume subsequent negative swabs must be false negatives.

Patients colonised and infected with MRSA were treated as equally likely to transmit, in order to run a simple, two-state model. Differentiating between the two cases would require additional data on the onset of clinical symptoms for infection, as well as parametrisation of the transition from the colonised to infected states. It has been shown that there is not a significant difference in transmission between these groups [176].

Only routine pooled screening results were used to evaluate the presence of MRSA collected at the start and end of a patient's episode. It is certainly possible for patients to be colonised at one site but not another [9], and patients who are colonised or infected at a site other than the screening sites, such as wounds and surgical sites, may potentially be missed. The colonisation pressure may be underestimated by considering only colonisation at these sites. In the next section, we investigate the incorporation of additional screening results from clinical sites.

Compliance with contact barrier precautions or the use of decolonisation treatment was not assessed. In reality, a healthcare worker may be reminded to adhere to precautions by having to enter a side room. For the same reason, compliance with decolonisation therapy may be increased. These factors could potentially result in a benefit to side room isolation over open ward isolation under our model assumptions. Any additional risk reduction from this psychological effect is incorporated into our estimates of E_{iso}.

We are not aware of a similar study estimating the effectiveness of isolation and decolonisation treatment in hospital general wards, so directly comparable results are not

available. A model based analysis by Kypraios et al. was designed to assess the impact of isolation on MRSA transmission, using data collected from eight ICUs in a Boston hospital [52]. Isolation in this study was considered to be barrier precautions; that is, the wearing of gowns and gloves. All ICU beds in this setting were in single rooms. Bayesian inference was used to derive estimates for unisolated positive days and the probability of isolation effectiveness. In this study, isolation effectiveness was assessed with the measure a_1/a_2 (which approximates our measure E_{iso}), and this was estimated to be 0.75 (95% CI: 0.25, 2.22), pooled across each of the ICUs.

The effect of decolonisation therapy on MRSA transmission rates is unclear. A systematic review on the effect of mupirocin nasal ointment on *S. aureus* infection rates in nasal carriers found evidence to support its effectiveness, but reduction in transmission was not investigated in any included study, and only one considered methicillin-resistant strains separately [177].

2.6.3 Pseudolikelihood approach

The pseudolikelihood analysis was significantly faster than the MCMC approach, and provides crude parameter estimates. The pseudolikelihood is calculated under additional assumptions, compared to the Bayesian approach, namely, that patient episodes are independent, and that colonisation pressure is fixed. The colonisation pressure here is likely to be underestimated, since undetected carriage is not taken into account. We find that estimates for p and z compare quite well to the values attained in the Bayesian analysis (figure 2.9), but transmission parameters a_0 and a_2 were often estimated to be very close to zero.

In order to estimate the isolation effectiveness E_{iso} with the pseudolikelihood, we calculated the median of the statistic from 1000 bootstrap samples. This is relatively time-consuming, and performing this eliminates any time-saving advantages of the pseudolikelihood approach over the MCMC algorithm.

It appears that this approach does not offer many advantages over Bayesian methods, and requires more restrictive assumptions. It may be useful to generate quick and crude estimates of the parameters. These estimates could potentially be used as starting points for the MCMC algorithm, and the estimated variance might be used to initially calibrate the proposal distribution for the Metropolis-Hastings step.

2.6.4 Summary

Our analysis indicates isolation in combination with decolonisation treatment is associated with a reduction in MRSA transmission of around 64% in hospital general wards, and that approximately three-quarters of ward transmission is due to unisolated colonised patients. We estimated that over 50% of colonised patient-days were spent out of isolation. Therefore attempting to minimise this figure would be key to reducing MRSA acquisitions in this setting. Further research into the separate effects of decolonisation treatment and isolation, as well as a more informative estimation for the difference between open ward and side room isolation would be required to provide statistical evidence supporting the precise components of a package of interventions for a newly-discovered positive patient. With appropriate data from a purpose-designed trial, our model may be extended to consider these factors.

2.7 Clinical isolate data

In the previous analysis, one of our assumptions was that a patient with a 'colonised' status was colonised at one of the screening sites (nose, axillae, groin), and that this was a sufficient measure to determine MRSA carriage. In reality, an individual may be colonised at other sites, particularly in the case where a patient is intubated or has an open wound. Such a patient may be colonised (or infected) at this site, and yet remain negative at the screening sites, resulting in a true negative swab result. Using the surveillance screening data alone may therefore ignore a potentially significant source of transmission.

Using additional positive clinical isolate results taken from high-risk patients or suspected MRSA carriers during the clinical trial at GST, we investigate the validity of using surveillance screening alone, and the effect of incorporating these additional data on the transmission parameters.

2.7.1 Data

In the study by Jeyaratnam et al. [50], patients were screened at the nose, axillae or groin (which we define as 'screening sites') on admission and discharge, as well as at skin breaks or clinically indicated sites ('clinical isolates') if applicable. All positive (but not negative) clinical isolate results were recorded, but were not used in our initial

Admission-discharge swab pairs

A

Admission \ Discharge	⊕	⊖	NA	
⊕	316	134	189	639
⊖	259	7520	4564	12343
NA	25	421	599	1045
	600	8075	5352	14027

B

Admission \ Discharge	⊕	⊖	NA	
⊕	125	75	247	447
⊖	184	7388	4622	12194
NA	9	411	966	1396
	318	7874	5835	14027

C

Admission \ Discharge	⊕	⊖	NA	
⊕	595	75	153	823
⊖	296	7388	4510	12194
NA	19	411	580	1010
	910	7874	5243	14027

Table 2.9: Total admission-discharge swab pairs under different assumptions. Positive swabs are indicated by ⊕, while ⊖ denotes negative results. In all settings, 'NA' covers missing swabs, or those which are invalid under the conditions described in Jeyaratnam et al. **A)** Total swab pairs using only surveillance swabs on admission and discharge. **B)** Swab pairs using surveillance swabs only, having deleted results occurring after a clinical positive. This 'override' method is described below. **C)** Swab pairs using surveillance swabs and positive clinical isolates. A patient is considered positive if they have received a positive result from any body site.

analysis of MRSA transmission. The work by Jeyaratnam et al. assumed patients to be positive on admission to a study ward if they were screened positive via any specimen taken up to five days prior to hospital admission, or 48 hours after ward admission, or were transferred from another hospital where they were screened positive. Tables 2.9a and 2.9c show the differences between using screening swabs only versus additional data. We see almost a 30% increase in patients observed to be positive on admission. This indicates that a large amount of MRSA carriage is ignored when considering only

surveillance swabs. Clearly, it is important to consider how these additional data affect our analysis of MRSA transmission. However, certain difficulties arise in attempting to incorporate the additional screens. While surveillance screens are taken routinely for all patients, clinical isolates are taken only from high risk patients, suspected to be MRSA carriers. Simply treating these results in the same way would introduce bias, particularly as negative clinical isolates are not reported. Further, the location of colonisation becomes important. Once a patient is known to be colonised at a wound site, there is no reason to assume that a subsequent negative swab at the screening sites is a false negative, and no information to suggest how likely it is for MRSA to be transferred from one body site to another. In fact, the notion of treating patients as simply 'colonised' or 'susceptible' becomes questionable. As such, we now explore different methods to incorporate this additional data.

2.7.2 Methods

We consider the situation where patients either have a pair of surveillance swabs alone, or these plus additional clinical isolates.

- **Treat both sets of results equally.** As previously mentioned, the methods and reasons for collection of the two sets of results differ, meaning that treating them identically will likely introduce important biases. Firstly, since colonisation can occur at a single site, the relation between the result of two swabs taken at different sites is complex, and a simple two-state Markov chain of patient status progression cannot suitably accommodate this. Secondly, using the same measure of sensitivity for these results is likely to be unreasonable, due to the different methods of collection. In any case, since no negative clinical isolate results are recorded, there is no way to estimate the sensitivity of the clinical isolate.

  To deal with the sensitivity issue, we could redefine the quantity z as

  $$z = P(\text{screened positive} \mid \text{positive at any site}),$$

  which is no longer a measure of sensitivity, and not a greatly informative or clinically interesting quantity. Alternatively, we could remove the parameter z altogether, and assume 100% sensitivity. In this case, we would need a method to deal with a negative result following a positive result. The simplest solution would be to exclude these results.

- **Override surveillance swabs with clinical isolates.** This method would override any surveillance swabs taken after the date of any positive clinical isolate. Once this has occurred, the patient is treated as positive and subsequent surveillance swabs are ignored. In particular, they are not considered in the calculation of sensitivity. This means that the remaining surveillance swabs can be used to fairly assess the sensitivity, and an individual is regarded as MRSA positive if they are colonised at either the screening sites or elsewhere. Unfortunately, this involves disregarding a large portion of the surveillance swabs, which will increase uncertainty around the estimate of sensitivity. Since those with a clinical positive screen are at a greater risk of colonisation at a screening site, we lose a significant proportion of positive swabs in particular. Since some wards had a very low MRSA prevalence during the study, the information remaining to inform the estimate of sensitivity is minimal.

- **Introduce more patient states.** We could expand the model beyond the current, two-state model by allowing for colonisation at different sites. This allows patients with a positive clinical isolate to progress to a different state than those who are positive at a screening site. This also allows patients with a positive wound swab to be legitimately screened negative at the screening sites. We can formulate different transmission rates to look at progression between the multiple colonisation statuses (see figure 2.10). The drawback of this approach is its considerably greater complexity than the current framework. The probability of becoming colonised at a site other than the screening sites is heavily dependent on whether the patient has a wound, or is intubated, and complete data regarding a patient's circumstances in this respect would be required. Even with these data, since colonisation is relatively infrequent, there still may not be enough MRSA positive patients to separate the 'colonised' state into multiple states and derive meaningful estimates for transmission parameters. One would have to take into account the fact that clinical isolates are not collected in the same manner as surveillance swabs, and are taken from a selected subpopulation. Moreover, an assumption would have to be made about the sensitivity of the clinical isolate swabs.

All the above methods either introduce bias, ignore data, or are potentially too complex to derive meaningful results. We decided that the best way to proceed with analysis would be to use the second method, overriding surveillance swabs with positive clin-

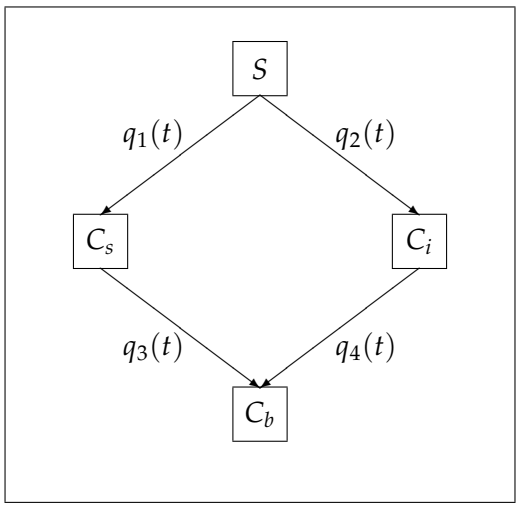

Figure 2.10: Extended four-state model. This allows patients to be colonised at the screening sites (C_s), at another site, determined by a clinical isolate,(C_i), or both (C_b). Up to four different transmission rates can be associated with this model ($q_.(t)$). We would also require a parameter for the probability of being admitted in each state.

ical isolate results. This should not unduly affect any of the parameters, and should more accurately reflect the burden of MRSA in the ward. Note that estimates of sensitivity are informed by the swab results in table 2.9b.

A crude estimate of sensitivity can be obtained by dividing the number of positive-positive pairs by the total number of patients with a positive admission swab and a valid discharge swab. This uses the assumption that positive patients remain positive, and that specificity is 100%, and so treats a negative result following a positive screen as a false negative. This is the minimum number of false negatives; the actual number may be higher, and is inferred with the data augmentation algorithm. Table 2.10 presents a summary of the data informing the estimation of z under the override assumption. The upper bound for the parameter z is calculated as

$$1 - \frac{\text{Number of positive-negative pairs}}{\text{Number of positive swabs}}.$$

It is clear from table 2.10 that the information available to estimate z varies considerably across the ten study wards, and is fairly minimal for some settings. Notably, ward 9 has only 22 positive swabs, and ward 5 has a crude estimate of sensitivity of just 0.2.

We repeated the analysis using the same methods as the previous analysis, using a data-augmented MCMC algorithm to allow a patient's colonisation time to vary between admission time, and time of any first positive result (or discharge when no

Data available to inform sensitivity estimate

Ward	1	2	3	4	5	6	7	8	9	10
$\oplus \to \oplus$	6	10	33	27	3	12	11	9	3	11
$\oplus \to \ominus$	6	7	18	6	12	7	5	7	3	4
Total positive swabs	48	88	175	124	47	82	71	57	22	51
Crude estimate z	0.50	0.59	0.65	0.82	0.20	0.63	0.69	0.56	0.50	0.7
Upper bound z	0.89	0.93	0.91	0.95	0.8	0.92	0.93	0.89	0.88	0.93

Table 2.10: Analysis of remaining surveillance swabs for each ward, having removed those overridden with a positive clinical specimen. The top row is the number of patients who received a positive screen ($\oplus$) on admission and discharge, and the second row is the total number of patients with a positive admission swab, and a negative discharge swab ($\ominus$). We can derive an upper bound for sensitivity from the fixed total number of (true) positive swabs, and minimum number of false negative swabs (total number of $(\oplus, \ominus)$ pairs).

positive results), but surveillance swabs occurring after clinical positive swabs were discarded.

2.7.3 Results

As anticipated, the prevalence on admission, p, is estimated to be higher when using the additional clinical results. A pooled estimate of this parameter is 0.09 (95% CI: 0.07, 0.11), compared to 0.07 in the original analysis. The posterior mean of sensitivity, z, is lower, and the variance quite a lot higher than previously. The sensitivity, z, was estimated to be 0.57 (0.42, 0.71). The validity of this estimate is discussed later. We found that the transmission parameters differed from the initial estimates; the unisolated colonised parameter (a_1) increased slightly, while the isolated and colonised parameter (a_2) had decreased. As a consequence, the estimate of isolation effectiveness is somewhat higher; we estimate E_{iso} to be 0.23 (0.13, 0.42), which corresponds to a reduction in transmission of 77% (58%, 87%), compared to the 64% previously. The estimates are summarised in table 2.11.

We found that the data available for estimating sensitivity was limited, and the uncertainty surrounding this parameter was high. For this reason we decided to investigate the use of an informative prior for this parameter. A Beta(12,3) distribution was chosen, which corresponds to a mean of 0.8 and a variance of 0.01. This is a reasonable prior

estimate, given values found in the literature [138] and our own estimate of 0.77 from the previous analysis.

Using this informative prior increased our posterior estimate of sensitivity to 0.66, and reduced the estimate of p slightly, to 0.08. However, the effect of isolation remained almost the same at 0.22 (0.12, 0.4).

Parameter estimates under varying prior sensitivity means

Prior mean	Posterior estimates					
z	p	z	a_0	a_1	a_2	E_{iso}
U*	0.091	0.57	2.6×10^{-4}	20.8×10^{-4}	1.90×10^{-4}	0.23
0.6	0.086	0.60	2.5×10^{-4}	21.0×10^{-4}	1.83×10^{-4}	0.22
0.7	0.085	0.63	3.1×10^{-4}	21.0×10^{-4}	1.82×10^{-4}	0.25
0.8	0.084	0.66	2.6×10^{-4}	20.7×10^{-4}	1.80×10^{-4}	0.22
0.9	0.084	0.65	2.6×10^{-4}	20.8×10^{-4}	1.87×10^{-4}	0.23

Table 2.11: Posterior parameter estimates under different prior assumptions. Sensitivity, z, takes a Beta distribution with a mean varying between 0.6 and 0.9, and variance 0.01. The table gives posterior means for the parameters θ, and the median for E_{iso}. U* indicates that a uniform prior distribution on (0,1) was used for z.

Table 2.11 shows how parameter estimates change as we alter our prior assumptions about sensitivity. The estimates are quite robust to the choice of sensitivity prior, and the estimate for isolation effectiveness remains high under all prior means. In addition, we ran an even more informative prior, Beta(127.8, 32.1), corresponding to a prior mean and variance of 0.8 and 0.001 respectively. The prior information for z was such that the posterior distribution remained similar to the prior for most wards. The pooled posterior sensitivity was 0.78 (0.76, 0.80), and E_{iso} was 0.23 (0.15, 0.37). This seems to confirm that this estimate does not depend on the prior of z.

For comparison, we also ran the analysis under the assumption that all swabs were the same. As mentioned previously, this will unfairly increase the estimate of z, but we might ignore this to see the effect this has on our main interest, the estimate of isolation effectiveness. The results are summarised in table 2.12.

Pooled posterior estimates, treating all swabs equally

Parameter	Estimate
p	0.07 (0.05, 0.09)
z	0.92 (0.89, 0.94)
a_0	2.6 (0.9, 4.2)$\times 10^{-4}$
a_1	22.7 (16.3, 29.1)$\times 10^{-4}$
a_2	1.7 (0.6, 2.7)$\times 10^{-4}$
E_{iso}	0.20 (0.11, 0.35)

Table 2.12: Pooled parameter estimates, treating clinical isolates and surveillance swabs as the same. This corresponds to table 2.9c.

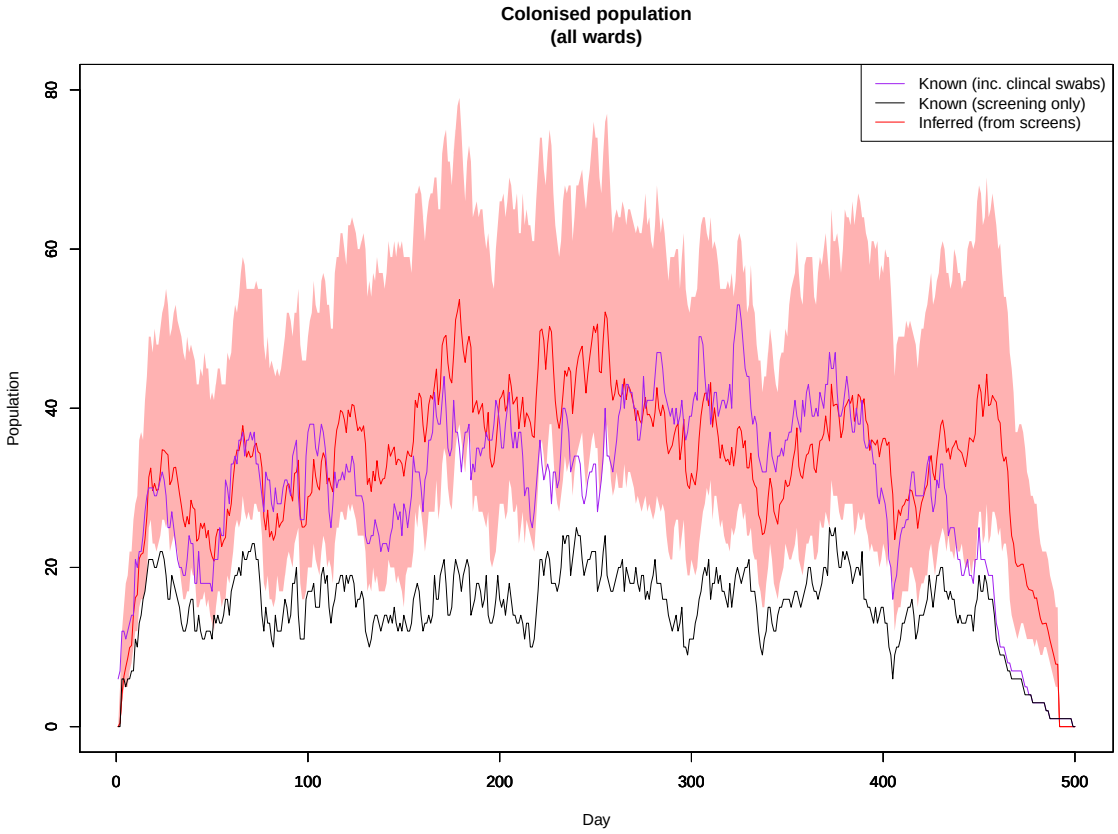

Figure 2.11: Inferred and observed colonised population. We plot the population count of patients observed to be positive by the screening only (black), as well as the number observed to be positive via either a screening test or a clinical isolate (purple). The colonised population inferred from the surveillance screens is shown in red (95% credible interval shaded in pink).

2.7.4 Discussion

Repeating this analysis with additional clinical isolate data has revealed additional colonisation pressure that was otherwise undetected by surveillance swabs alone. The majority of 'extra' patients will have been in isolation, as hospital staff were aware of their colonisation/infection status. This means that the number of colonised patients in isolation ($C_I(t)$) is greater than previously thought, while the transmission rate is similar or lower than we had previously estimated. The result of this is that the parameter a_2 has decreased, while a_1 is increased to accommodate the fewer inferred colonised patients out of isolation. This results in a similar transmission rate, but a higher estimate of isolation effectiveness.

The initial analysis is believed to be a slightly conservative estimate of prevalence and transmission, given that many MRSA positive patients are undetected by using only surveillance swabs, and these patients are most likely to contribute to the colonised and isolated population count. In this investigation, we have shown that by including the positive clinical isolates, the effect of isolation and decolonisation treatment is higher. It was estimated that while few positive patients are unisolated, these patients have a much larger effect on the transmission rate. However, we have acknowledged that the methods used to incorporate this additional data are imperfect, and may introduce uncertainty and bias.

With only positive clinical isolate results available, it is not possible to estimate the sensitivity of this set of swabs, which must therefore be treated the same as the screening results, or used only to inform the colonised population count. Due to the nature of the additional data, we could find no method to fully incorporate this information without introducing bias or having to ignore parts of the dataset. We found that overriding the surveillance data with clinical isolates removed a large proportion of positive swabs, which are required to inform the sensitivity estimate. This meant that the estimate of sensitivity was much more uncertain than previously. We found that lower sensitivity values than we might expect are estimated, unless a very informative prior distribution is used for z.

Incorporating this additional data has given us a better idea of the true prevalence of MRSA at any given time. We compared the total observed MRSA prevalence using all available swab data to the mean inferred colonised population estimated from the MCMC algorithm in figure 2.11. The estimated prevalence, derived using only the surveillance screening results approximates the total observed prevalence reasonably

well, providing some confidence in our original method.

We have shown how the additional burden of MRSA positive patients affects our parameter estimates, under certain conditions. We consider the initial analysis to be more reliable, but this additional investigation supports our claim that isolation and decolonisation treatment reduce the MRSA transmission rate in this setting.

CHAPTER 3

Bayesian model selection

3.1 Introduction

In statistical modelling, one is often faced with a system that may be modelled in several different ways. Having derived estimates from a particular model, it is important to consider how plausible this model is, relative to other alternative formulations. Model selection might be performed to either select a 'best' model (eg. [59]), to investigate the dynamics generating observed data [178], or to derive weightings for a model averaging procedure [179]. Many studies on healthcare-associated infections have derived estimates for transmission parameters (eg. [145, 149]), however, few have attempted to compare alternative underlying models. While models are often constructed with the aim of providing answers to specific questions (such as estimating the effectiveness of isolation in the previous chapter), there may well be a model which describes the fundamental dynamics of the observed (or indirectly observed) system better. It is of interest to consider the support for a range of alternatives to see how much, if at all, the data support the use of a particular model.

While the Bayesian methods described in the previous chapter provide a convenient and flexible framework in which to derive parameter estimates, the procedures to evaluate and compare models in this setting are not as universally accepted or as straightforward as those in a classical framework. While the Bayes factor is a well-known tool which may be used to compare pairs of (not necessarily nested) models in a Bayesian framework, computation can be difficult, and there is a dependency on the specification of the prior distributions which can be troublesome, especially for models of differing dimension, as we discuss later. Furthermore, models which account for missing data introduce additional difficulty in the definition of model complexity.

Much research has been undertaken in the past decades to develop sophisticated tools to evaluate Bayesian models, and while often computationally expensive, these have been shown to perform well. There are, however, few studies evaluating systematically the performances of Bayesian model selection methods, and for the particular case of epidemic models, a lack of analyses which consider multiple model comparison techniques. In this chapter, we propose to compare a set of transmission models, in order to assess the performance of two important Bayesian model selection methods: reversible jump Markov chain Monte Carlo (RJMCMC), and the deviance information criterion (DIC). We perform a systematic study using both real and simulated data to investigate the effect of several factors on the outcome of the two methods.

In section 3.2 of this chapter, we provide an overview of Bayesian model choice methods, firstly discussing the calculation and interpretation of Bayes factors. We describe the difficulty of estimating the marginal likelihood, key to finding estimates of posterior model probabilities. We then introduce the DIC and RJMCMC, as well as alternative MCMC-based model choice methods, and lastly describe an ABC approach.

Section 3.3 describes existing analyses of epidemics which have used Bayesian model selection methods. We discuss the additional complexities which arise in the case of models with missing data, and the particular case of analysing transmission of hospital pathogens. We then define the models which we intend to compare in our study.

The main components of this chapter are detailed assessments of the performance of two widely-used Bayesian model choice approaches, RJMCMC and the DIC. In section 3.4 we consider the application of the RJMCMC algorithm to a series of transmission models, with the aim of calculating posterior model probabilities. We then conduct a series of simulation studies, exploring some of the issues which affect its performance in this setting. These include the specification of prior distributions, the choice of transformation function allowing between-model jumps, and the true transmission rates of the system. We report our findings, and then estimate posterior model probabilities using real hospital data, examining the results in light of the findings from our simulation studies.

In section 3.5, we investigate a version of DIC for missing data models, evaluating its performance in choosing between a set of transmission models. Once again, we perform a systematic analysis, determining where the DIC can identify the correct model under various simulated datasets. We finally calculate the DIC for each transmission model, using real hospital data.

Section 3.6 concludes the chapter with a discussion of our results for both approaches. We discuss the advantages and disadvantages of both methods, and under which conditions we might expect reasonable results.

3.2 Background

3.2.1 Bayesian model selection methods

Suppose we have data x, which we assume have been generated under one of a set of k candidate models, $\mathcal{M} = \{m_1, \ldots, m_k\}$. We suppose that m is an indicator of the model, which is characterised by a set of parameters $\theta_m \in \Theta_m$. Of fundamental interest in Bayesian model selection is the evaluation of the posterior model probability

$$\pi(m|x) \propto \int_{\Theta_m} \pi(x|\theta_m, m)\pi(\theta_m|m)\mathrm{d}\theta_m \pi(m), \tag{3.2.1}$$

where $\pi(x|\theta_m, m)$ is the likelihood of the data under model m, $\pi(\theta_m|m)$ is the prior density of parameters θ_m for model m, and $\pi(m)$ is the prior probability of model m. Evaluation of the posterior model probability is often achieved via within and between model sampling using MCMC-based methods. Such approaches aim to provide a relative measure of fit, and quantify evidence in favour of a particular model, rather than providing a 'best' model. Alternatively, information criteria provide models with a score, based on the relative model fit (how closely the model fits the observed data compared to competing models), and penalising by model complexity (often considered equivalent to the number of parameters in the model). Such measures are more commonly used to select one model from a set of candidates, and do not provide a straightforward way to assess relative fit of models.

3.2.2 Bayes factors and the marginal likelihood

The Bayes factor may be used to assess the evidence in favour of one model over another, conditional on one of the models being true. The Bayes factor has been widely used for Bayesian hypothesis testing and model comparison, despite complexities of calculation and its dependency on the choice of model-specific prior distributions [133–135].

The Bayes factor of models l and m is the ratio of marginal likelihoods,

$$BF(l, m) = \frac{\pi(x|l)}{\pi(x|m)} = \frac{\int \pi(x|\theta_l, l)\pi(\theta_l|l)\mathrm{d}\theta_l}{\int \pi(x|\theta_m, m)\pi(\theta_m|m)\mathrm{d}\theta_m}. \tag{3.2.2}$$

The Bayes factor is often used for model selection purposes, although has been criticised for its dependence on the prior distributions of the model parameters. Lindley discussed the paradoxical support exhibited by posterior odds for models with an informative prior distribution over models with a diffuse prior, in the case of point hypothesis testing [180]. In many cases, a parameter's prior distribution is desired to have as little impact on the posterior estimate as possible. However, as a prior distribution becomes more diffuse, the preference shown by the Bayes factor for the alternative model becomes larger, due to the smaller expected likelihood. This has become known as Lindley's paradox. Comparing models of different dimensions is of particular interest in our study. If uninformative prior distributions are assigned to the model parameters, increasing dimensionality will tend to reduce the expected likelihood under the priors, and the same effect is observed — that is, the Bayes factor will support the model with fewer parameters.

Since direct calculation of the Bayes factor requires the calculation of the marginal likelihood, which is commonly analytically intractable, an alternative method is required. Numerical approximation methods have been suggested to derive estimates of the marginal likelihood $\pi(x|m)$. Chib proposed methods to approximate the marginal likelihood using a Gibbs sampler, relying on the availability of full conditional distributions in closed form [181]. Since the marginal likelihood can be expressed as

$$\pi(x|m) = \frac{\pi(x|\theta_m, m)\pi(\theta_m|m)}{\pi(\theta_m|x, m)},$$

it suffices to evaluate the likelihood and prior distribution at a point θ^*, as well as provide an estimate of the posterior density at θ^*, $\hat{\pi}(\theta^*|x, m)$. By partitioning the parameter vector into blocks $\theta = \{\theta_1, \ldots, \theta_B\}$ which may be updated using a Gibbs step, the posterior density may be written as the product

$$\hat{\pi}(\theta^*|x, m) = \prod_{i=2}^{B} \hat{\pi}(\theta_i^*|\theta_{i-1}^*, \ldots, \theta_1^*, x, m).$$

Now, each component may be estimated as

$$\hat{\pi}(\theta_i^*|\theta_{i-1}^*, \ldots, \theta_1^*, x, m) = \frac{1}{N}\sum_{j=1}^{N} \pi(\theta_i^*|\theta_1^*, \ldots, \theta_{i-1}^*, \theta_{i+1}^{(j)}, \ldots, \theta_B^{(j)}, x, m),$$

where $\theta_{i+1}^{(j)}, \ldots, \theta_B^{(j)}$ are the jth samples from a Gibbs algorithm, using the full conditional density with $\theta_1^*, \ldots, \theta_{i-1}^*$ fixed. This approach was later extended to estimation of the marginal likelihood, based on Metropolis-Hastings output [182].

Friel and Pettitt proposed a method to estimate the marginal likelihood using ideas from path sampling [134]. The authors consider the 'power posterior', $\pi_t(\theta|x, m) \propto$

$\pi(x|\theta,m)^t\pi(\theta|m)$; by allowing t to move from 0 to 1, the power posterior follows a path from the prior to the posterior density. The authors show that

$$\log(\pi(x|m)) = \int_0^1 E_{\theta|x,t,m}(\log(\pi(x|\theta,m)))\mathrm{d}t,$$

where the integrated term is the expectation of the log likelihood, with respect to the power posterior, with power t. Partitioning t on the unit interval allows this to be approximated by numerical integration, and the expectation may be estimated by drawing samples from $\pi_t(\theta|x)$ for particular levels of t.

3.2.3 RJMCMC

3.2.3.1 Description

Rather than viewing competing models and their corresponding parameter sets separately, it is convenient to consider a joint parameter space, spanning each model $m \in \mathcal{M}$ and its individual parameter space $\theta_m \in \Theta_m$. The RJMCMC algorithm, proposed by Green in 1995 [124], generates a chain of points $(m^{(1)},\theta^{(1)}),\ldots,(m^{(N)},\theta^{(N)})$, where $\theta^{(i)} \in \Theta_{m^{(i)}}$, which converges to the posterior distribution $\pi(m,\theta|x)$. Points are sampled across the space $\mathcal{M} = \bigcup_{m=1}^{k}\{m\}\times\Theta_m$, where proposal points are dependent on the current state, allowing for moves between models of differing dimension.

The RJMCMC algorithm may be implemented in the following way:

Reversible jump MCMC algorithm

1. Set initial model $m^{(0)}$, parameters $\theta^{(0)}$ and number of iterations N.

2. Let $(m^{(i)}, \theta^{(i)})$ indicate the current state. With probability $j(m^{(i)}, m^*)$, propose a move from the current model $m^{(i)}$ to model m^*.

3. If $m^{(i)} = m^*$, propose parameters $\theta^* \in \Theta_{m^*}$ according to a within model Metropolis-Hastings step. Otherwise, sample a random vector $u \sim h_{m^{(i)},m^*}$, and propose the candidate parameter vector

$$g_{m^{(i)},m^*}(\theta^{(i)}, u) = (\theta^*, u').$$

4. Set $(m^{(i+1)}, \theta^{(i+1)}) = (m^*, \theta^*)$ with probability $\min(1, \alpha)$, where

$$\alpha = \frac{\pi(x|m^*, \theta^*)\pi(\theta^*|m^*)\pi(m^*)h_{m^*,m^{(i)}}(u')j(m^*, m^{(i)})}{\pi(x|m^{(i)}, \theta^{(i)})\pi(\theta^{(i)}|m^{(i)})\pi(m^{(i)})h_{m^{(i)},m^*}(u)j(m^{(i)}, m^*)}|J|,$$

and

$$J = \left| \frac{\partial g_{m^{(i)},m^*}(\theta^{(i)}, u)}{\partial(\theta^{(i)}, u)} \right|,$$

otherwise $(m^{(i+1)}, \theta^{(i+1)}) = (m^{(i)}, \theta^{(i)})$.

5. If $i < N$, go to step 2.

This is a generalisation of the within-model Metropolis-Hastings algorithm (described in section 1.5.2.3) to sample over a parameter space spanning all models. The number of possible models is not required to be finite, and the algorithm may be implemented without knowledge of the size of the model [183].

RJMCMC requires mechanisms to move between models and their associated parameters. In the above algorithm, the function $g_{m^{(i)},m^*}(u, m^{(i)}, \theta^{(i)}) = (\theta^*, u')$ is a deterministic diffeomorphism which transforms parameters between models, given the current state and a random variable $u \sim h_{m^{(i)},m^*}$. This is constructed such that

$$\dim(u) + \dim(\theta_{m^{(i)}}) = \dim(u') + \dim(\theta_{m^*}).$$

Green showed that the RJMCMC algorithm generates a Markov chain satisfying the detailed balance condition, and which converges to the posterior distribution $\pi(m, \theta|x)$ [124]. One may estimate the posterior model probability, $\pi(m|x)$, as the proportion of accepted points in model m; that is, $\pi(j|x) \approx \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(m^{(i)} = j)$. Furthermore, within-model parameter estimates for a given model j may be derived from the sample $\{\theta^{(i)}_{m^{(i)}}|m^{(i)} = j\}$, provided that this set is of a reasonable size. It is clear that the behaviour of the algorithm depends on the choice of transformation function, g, the random variable proposal density h, as well as prior information.

In jumping from model m to l, a sensible choice of transformation, g, would map the current point θ_m to an 'equivalent' point θ_l in the parameter space of l, such that the proposed point has similar posterior support as the current state [184]. Proposing a point away from the region of high posterior probability density in a candidate model will very likely result in rejection [185]. This may be avoided by using some form of summary function $S(\cdot)$ which is applicable to all models, and choosing θ_l such that $S(\theta_l) = S(\theta_m)$. In the setting of epidemic models, where interest often lies in fitting various forms of transmission rate, matching these rates in a between-model move might be attempted, by selecting appropriate candidate parameter values. This increases the chance that the proposed point will be accepted.

3.2.3.2 Transdimensional independence sampler

If the within-model posterior distribution of the transmission parameters is known, or can be fairly well approximated, it may be more efficient to sample from this distribution directly, independent of the current state of the algorithm. Although it is normally difficult to approximate the posterior distribution of all within model parameters *a priori*, by doing so, one can avoid having to specify an efficient proposal mechanism, which can often be problematic.

We assume that $\theta_i|x \sim h_i$ for all $i \in \mathcal{M}$, the set of all models. Given the current state is $(m^{(i)}, \theta^{(i)})$, where $\theta^{(i)} \in \Theta_{m^{(i)}}$, a move to model m^* is proposed with probability $j(m^{(i)}, m^*)$, and a random vector u is drawn from h_{m^*}. The transformation function is given as $g_{m^{(i)},m^*}(\theta^{(i)}, u) = (v, u) = (v, \theta^*)$, where $\dim(v) = \dim(\theta^{(i)})$. This move is independent of $\theta^{(i)}$, and the Jacobian will be equal to 1. The proposed point is now accepted with probability $\min(1, \alpha)$, where

$$\frac{\pi(x|m^*, \theta^*)\pi(\theta^*|m^*)\pi(m^*)j(m^*, m^{(i)})h_{m^{(i)}}(\theta^{(i)})}{\pi(x|m^{(i)}, \theta^{(i)})\pi(\theta^{(i)}|m^{(i)})\pi(m^{(i)})j(m^{(i)}, m^*)h_{m^*}(\theta^*)}.$$

Clearly, the independence sampler is highly sensitive to the choice of proposal. The acceptance rate will depend on how closely the proposal relates to the posterior, and a diffuse proposal distribution will lead to extremely low acceptance. Since little is usually known about the form of the posterior, the independence sampler will be a very inefficient method of deriving estimates without pilot investigations into the within-model posteriors. One possibility to improve the efficiency of this process would be to run within-model MCMC algorithms to derive posterior distributions for each $\pi(\theta_m|m,x)$, and use these to inform the choice of proposal distribution. If the number of potential models is large, running separate within-model analyses for each will also be time-consuming.

3.2.4 DIC

3.2.4.1 Description

Information criteria are statistics which measure the relative adequacy of a given model. This typically involves a trade-off between parsimony and goodness-of-fit, the aim being to select a simple model which fits well. An example of this is Akaike's information Criterion (AIC), defined as

$$AIC = 2\nu - 2\log L(\hat{\theta}),$$

where ν is the number of parameters in the model, and $\hat{\theta}$ is the maximum likelihood estimate of the likelihood L [186]. In a system of hierarchical models, a more complex model will always provide a better fit, and so the AIC measures model adequacy (given here by the maximum likelihood), penalised by complexity (number of parameters).

The relative fit of Bayesian models may be checked informally via the posterior deviance, an approach suggested in 1974 by Dempster [187]. The posterior deviance is defined as

$$D_x(\theta) = -2\log(\pi(x|\theta)).$$

The expected deviance, $\overline{D_x(\theta)} = E_{\theta|x}(D_x(\theta))$, takes lower values when values of θ result in a higher posterior probability [188]. The expected deviance may be estimated from samples of the posterior distribution;

$$\overline{D_x(\theta)} \approx \frac{1}{m}\sum_{i=1}^{m} D_x(\theta^{(i)}),$$

where $\theta^{(1)},\ldots,\theta^{(m)}$ are posterior samples, which may be taken from the output of an MCMC algorithm.

Such a measure can therefore be used to informally compare two different models. However, this alone does not take into account the relative complexity of the models, and there is no straightforward way to penalise a model for this, since the concept of model complexity is less clearly defined in a Bayesian setting. If we allow a prior distribution to become ever more informative, until there is a point mass on a particular value, a model can be viewed as decreasing in degrees of freedom, and therefore complexity, meaning that 'number of parameters', as used in the AIC, will not suffice.

Spiegelhalter et al. introduced the deviance information criterion (DIC) in 2002, as a generalised model comparison tool which can be used in a Bayesian setting, and is analogous to the AIC [136]. The authors describe the difficulty in defining complexity in a Bayesian setting, highlighting the 'level of focus' as a key issue in determining this. For instance, parameters relating to the specification of a prior distribution (hyperparameters) may or may not be of interest in an analysis, and their inclusion in the set of focussed parameters alters the complexity of the model. To contend with these issues, the authors proposed a complexity measure p_D, the effective number of parameters, which they define as

$$\begin{aligned} p_D &= \overline{D_x(\theta)} - D_x(\tilde{\theta}) \\ &= -2E_{\theta|x}\big[\log \pi(x|\theta)\big] + 2\log \pi(x|\tilde{\theta}), \end{aligned} \tag{3.2.3}$$

where $\tilde{\theta}$ is an estimator of θ, which was assumed in their paper to be the posterior mean. This is the difference between the average posterior deviance, and the deviance at the point $\tilde{\theta}$, which can be viewed as the degree of improvement (reduction in deviance) due to estimating $\tilde{\theta}$, over *a priori* knowledge. As prior uncertainty reduces, then one would expect this improvement to become smaller, corresponding to a reduction in complexity.

The DIC is then given as

$$\begin{aligned} \text{DIC} &= \overline{D_x(\theta)} + p_D \\ &= 2\overline{D_x(\theta)} - D_x(\tilde{\theta}) \\ &= -4E_{\theta|x}(\log \pi(x|\theta)) + 2\log \pi(x|\tilde{\theta}), \end{aligned} \tag{3.2.4}$$

the expected deviance plus the effective number of parameters, which, like the AIC, represents a trade-off between model fit and complexity.

Gelman et al. suggested an alternative formulation for p_D;

$$p_D = \frac{1}{2}\text{var}(D(\theta)|x),$$

where the variance of the deviance can be estimated from posterior samples [188].

3.2.4.2 DIC for models incorporating missing data

In cases where the deviance, or indeed, the likelihood $\pi(x|\theta)$, is not available in closed form, as is the case with missing data models (such as the MRSA transmission model with missing colonisation times, described in chapter 2), we cannot calculate the DIC as defined above. Suppose by introducing a set of unobserved data z, we may calculate the complete likelihood $\pi(x, z|\theta)$. In such cases, it is not immediately clear how to estimate the expected deviance $E_{\theta|x}D_x(\theta)$, due to the additional set of data z.

Celeux et al. discussed eight different formulations of the DIC for missing data models, extending the original definition of DIC for a wider range of models [137]. The authors discuss various interpretations of both $\tilde{\theta}$ and $\overline{D_x(\theta)}$, applicable when the likelihood cannot be explicitly calculated. The choice of these depends on the role of the missing data in the models to be compared; while in some cases they might be regarded as of no interest to the analysis in hand, the missing data themselves may in other cases, such as mixture models, be considered the focus of interest.

Consider the complete likelihood $\pi(x, z|\theta)$, where $\pi(x|\theta) = \int_z \pi(x, z|\theta)\mathrm{d}z$, and the joint deviance, $D_{x,z}(\theta) = -2\log\pi(x, z|\theta)$. We can consider the DIC for particular values of z to be

$$\begin{aligned}\mathrm{DIC}(z) &= 2\overline{D_{x,z}(\theta)} - D_{x,z}(\tilde{\theta}) \\ &= -4E_{\theta|z}(\log\pi(x, z|\theta)) + 2\log\pi(x, z|\tilde{\theta}).\end{aligned} \tag{3.2.5}$$

Now, by taking the expected posterior value of $\mathrm{DIC}(z)$ across the missing data z, we get

$$\begin{aligned}\mathrm{DIC}^* &= E_z(\mathrm{DIC}(z)) \\ &= E_z[-4E_{\theta|z}(\log\pi(x, z|\theta)) + 2\log\pi(x, z|\tilde{\theta})] \\ &= -4E_{z,\theta}(\log\pi(x, z|\theta)) + 2E_z(\log\pi(x, z|\tilde{\theta})),\end{aligned} \tag{3.2.6}$$

which, with $\tilde{\theta} = E_\theta(\theta|x)$, was described by Celeux et al. as DIC_6, which is appropriate in a setting such as this, where missing data are considered in order to make the complete likelihood tractable, but are not themselves the focus of interest. This measure performed adequately in their study, although the authors noted that it was possible to calculate negative values for p_D using this measure. The calculation of this measure is described below;

DIC for missing data models (DIC_6)

1. Run data-augmented MCMC procedure (see section 1.5.2.7) on the data x to draw samples from the posterior $\pi(\theta|x)$.

2. Derive the posterior mean $\tilde{\theta} = E_\theta(\theta|x)$ and the expected log-likelihood over θ and z;

$$E_{\theta,z}(\log \pi(x,z|\theta)).$$

3. Run data-augmented MCMC procedure a second time, this time with θ fixed at $\tilde{\theta}$, allowing only the augmented data to vary.

4. From the MCMC output, derive the expected log-likelihood over z, given $\tilde{\theta}$;

$$E_z(\log \pi(x,z|\tilde{\theta})).$$

5. DIC_6 is then given as

$$-4E_{\theta,z}(\log \pi(x,z|\theta)) + 2E_z(\log \pi(x,z|\tilde{\theta})).$$

3.2.5 Product space search

Carlin and Chib introduced the product space search to derive posterior model probabilities, based on an MCMC approach to sample across the space of all models. In contrast to the RJMCMC algorithm, which jumps between model parameter spaces of potentially varying dimension, the product space search explores the full, fixed dimension space $\mathcal{M} \times \prod_{j\in\mathcal{M}} \Theta_j$ [189]. The authors proposed a Gibbs step to sample from this space. This approach requires that the full conditional distribution for within-model parameters θ_j and model m are available, and makes the assumption that the data x are independent of $\{\theta_{j'}\}_{j'\neq j}$ under model j. The conditional distributions are then

$$\pi(\theta_j|\theta_{j'\neq j}, m, x) \propto \begin{cases} \pi(x|\theta_j, m=j)\pi(\theta_j|m=j) & m=j \\ \pi(\theta_j|m\neq j) & m\neq j, \end{cases}$$

where $\pi(\theta_j|m \neq j)$ are 'pseudopriors', and

$$\pi(m = j|\theta, X) = \frac{\pi(x|\theta_j, m = j)\pi(m = j)\prod_{l \in \mathcal{M}} \pi(\theta_l|m = j)}{\sum_{k \in \mathcal{M}} \left[\pi(x|\theta_j, m = j)\pi(m = j)\prod_{l \in \mathcal{M}} \pi(\theta_l|m = k)\right]}. \quad (3.2.7)$$

The posterior model probability for a given model j may then be estimated from the proportion of the output samples where $m = j$. The pseudopriors have no impact on the joint posterior $\pi(\theta_j, m = j|x)$, but are chosen to improve the efficiency of the algorithm. Performance is best when pseudopriors are close to the posterior [189]. However, having to draw from every pseudoprior at each iteration is a computational drawback for a large model space [190]. After a large number of iterations, the posterior model probability for model j may be approximated by calculating the proportion of samples $m^{(1)}, \ldots, m^{(n)}$ for which $m = j$. When comparing two models using the Bayes factor, the prior probabilities $\pi(m = j)$ may be chosen arbitrarily, since they do not affect this measure (see equation 3.2.2). This property means that they may be chosen to improve the mixing of the chain, allowing each model to be visited roughly equally [189]. This improves the efficiency in calculating the Bayes factor.

It is often not possible to sample from the full conditional distribution of parameters θ_j. Dellaportas et al. described a 'metropolised' version of the product space search, in which a Metropolis step is used to propose a model in each iteration, rather than drawing from the full conditional distribution $\pi(m = j|\theta, x)$ [191].

Godsill described a framework which allows parameters to be shared between models, generalising the product space search [192]. It can be shown that both Carlin and Chib's product space search and a version of RJMCMC are special cases within the framework.

3.2.6 ABC model choice

Toni et al. introduced a model choice procedure using approximate Bayesian computation (ABC) method, based on sequential Monte Carlo (SMC) [127]. The authors extend the within-model SMC algorithm [129] to include a model selection step. The within-model ABC SMC algorithm is described in section 1.5.2.9. As before, particles are sampled and passed through a series of filters, such that data simulated under particles from each generation approximate the observed data x ever more closely. For decreasing threshold values $\epsilon_1 > \cdots > \epsilon_T$ and a distance metric $\delta(\cdot, \cdot)$, particles in generation i represent a sample from $\pi(\theta|\delta(x, x^*) < \epsilon_i)$.

Prior to sampling a particle, a model is drawn according to the prior model probabil-

ities, $\pi(m)$. A model m particle is drawn from the set of weighted model m particles in the previous sample, and is perturbed. The proposed particle is used to simulate a new dataset x^*, and is accepted if the simulated data is close enough to the observed dataset. The final output of this algorithm is a set of weighted particles from various models, such that data simulated under these particles are sufficiently close the the observed data x. The proportion of final particles from a particular model represents an approximation to the posterior model probability. However, as pointed out by Robert et al., this estimate is highly dependent on the choice of $\delta(\cdot)$, even as ϵ_T approaches zero [193]. They point out that even if a comparison statistic is sufficient within models, it is not necessarily so between models, and an estimate of the Bayes factor using this approach will not necessarily converge to the true value.

3.2.7 Comparing model selection methods

There has been much debate surrounding the comparative performance of Bayesian model selection methods. Both Han and Carlin [190] and Dellaportas et al. [191] conducted studies into the performance of product space search methods and RJMCMC, comparing estimates of posterior model probabilities. Han and Carlin compared two linear regression models, in which exact posterior model probabilities had previously been calculated via numerical integration, and provided a target for the Monte Carlo methods. Dellaportas et al. compared a series of nested logistic regression models. Both studies found that estimates were very similar across all methods, although it has been suggested that the framework was not sufficiently complex to highlight poor performance in any method [184].

3.3 Bayesian model comparison for epidemic studies

3.3.1 Previous work

There are a few published examples of epidemic modelling in a Bayesian framework which have utilised model comparison techniques. A paper by Neal and Roberts in 2004 analysed data from a measles epidemic, and considered a series of nested models to describe the transmission dynamics [194]. The authors described a 'full' model, and four submodels, each of which was identical to the full model, with one parameter set to zero. A RJMCMC procedure was devised to move between these models in order

to explore the effect of school class and household spatial location on transmission. Reversible jump was also used by O'Neill and Marks in 2005 to consider whether vomiting episodes increased Norovirus transmission in a school [178]. Two models were compared, where one featured an additional parameter to account for the occurrence of vomiting. In both papers, simulation studies were performed to assess the impact of various assumptions made on the data.

Forrester and Pettitt conducted an analysis of MRSA transmission in ICUs, and used the DIC measure to compare nested models [58]. The definition of the DIC involves the calculation of the deviance of a model, which is typically not available in closed form for missing data models, leading to alternative formulations being proposed for such cases [137]. A later study by Cooper et al. in 2008 used an augmented data approach to model VRE transmission, and used an adaptation of the DIC for model comparison [126]. Since the posterior mean is not available over the missing dataset, the authors instead use the measure $p_D = \text{var}(D)/2$, an alternative measure of effective number of parameters [188].

Toni et al. introduced an ABC SMC model selection algorithm, and used SIR models to test its performance [127]. The authors created four competing deterministic models, and calculated posterior model probabilities for each model based on simulated data, attempting to identify the correct one. Their investigation into the performance of this method for SIR models was limited, but it was demonstrated that the basic model was identified correctly 664 times out of 1000 iterations. The authors note that larger simulated datasets result in stronger evidence for the correct model. The ABC SMC approach is sensitive to many factors, including choice of prior distributions and the choice of thresholds ϵ, which regulates which datasets simulated in the algorithm are close enough to the original dataset.

3.3.2 Aims of current work

In this study, we are interested primarily in using Bayesian model selection methods for the application to healthcare-associated infection data.

Throughout this chapter, three fairly simple and commonly used transmission models are compared, in order to determine how well model selection techniques can differen-

tiate between them. The models used are:

$$q_0(t) = a_0 \qquad (m{=}0)$$
$$q_1(t) = a_0 + a_1 C(t) \qquad (m{=}1)$$
$$q_2(t) = a_0 + a_1 C_N(t) + a_2 C_I(t). \qquad (m{=}2)$$

where $C(t)$ is the number of colonised patients at time t, of whom $C_N(t)$ and $C_I(t)$ are unisolated and isolated respectively. In this setting, model $m = 0$ assumes a constant transmission rate, independent of any colonisation pressure. Model 1 assumes that the transmission rate depends on the number of colonised patients present in the ward, but the effect from these individuals is the same no matter where they are located. Finally, model 2 separates the effect of isolated and unisolated colonised patients. In addition to transmission parameters, we estimate the probability of colonisation on admission, p, and the sensitivity of the MRSA carriage test, z, which are common to each model.

We work in discrete time, assuming that patients colonised on day t contribute to the colonised populations from day $t + 1$ until their discharge. Patients considered to be importations may transmit the pathogen from their day of admission.

As described in the previous chapter, we augment the model parameter space θ with the set of colonisation times and admission statuses A in order to achieve a tractable likelihood function. Within model m, we explore the posterior density

$$\pi(A, \theta_m | X) \propto \pi(X | A, \theta_m) \pi(A | \theta_m) \pi(\theta_m),$$

where

$$\pi(X | A, \theta_m) = z^{TP(X)} (1 - z)^{FN(X,A)},$$

and

$$\pi(A|\theta) = p^{\sum_i \phi_i} (1-p)^{n - \sum_i \phi_i} \cdot \prod_{i=1}^{n} \Big[\mathbf{1}_{t_i^c = t_i^a} + \mathbf{1}_{t_i^c \neq t_i^a} \exp\Big(- \sum_{t=t_i^a}^{\min(t_i^c - 1, t_i^d)} q_m(t) \Big) \Big] \prod_{\substack{j : t_j^c \neq \infty \\ \phi_j = 0}} (1 - e^{-q_m(t_j^c)}),$$

in which $TP(X)$ is the total number of true positive swab results, $FN(X, A)$ is the total number of false negative results, and ϕ_i is an indicator of carriage on admission for patient i. A full description of this model is given in section 2.4.

3.4 Reversible jump Markov chain Monte Carlo

3.4.1 Framework

A RJMCMC algorithm was used to derive posterior model probabilities for the MRSA transmission models described in the previous section. At each iteration, a move is proposed between these models at random, allowing the model dimension to change by one level of complexity, as shown in figure 3.1.

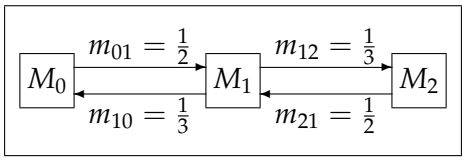

Figure 3.1: Model jump probabilities, where m_{ij} is the probability of jumping from model i to j.

From model 0 or model 2, a move to model 1 is made with probability $\frac{1}{2}$, otherwise no move is made. From model 1, a move is made to model 0 or 2 with probability $\frac{1}{3}$, otherwise no move is made. For each between-model move, candidate values must be resampled for the new parameter space, which may have increased or decreased in dimension, but p or z are not altered, since they are not directly affected by changing model. Similarly, the augmented data remains unchanged. Each model jump is now presented in detail.

Between models 0 and 1: This move increases the dimension of the parameter space by adding the parameter a_1. A transformation function is chosen such that the transmission rates for the current and proposed states are approximately equal. Candidate values $a_0^* = a_0 u$ and $a_1^* = a_0(1-u)/C_N$ are proposed, where u is a random variable drawn from h, which we define here to be the Uniform(0,1) distribution, and C_N is the average colonised population in the ward. This transformation is equivalent to the diffeomorphism $g : [0,\infty) \times [0,1] \to [0,\infty) \times [0,\infty)$, where

$$g(a_0, u) = (a_0 u, a_0(1-u)/C_N) = (a_0^*, a_1^*).$$

The Jacobian $|J|$ of the transformation associated with the move from model 0 to model

1 is

$$|J| = \left| \det \begin{bmatrix} u & a_0 \\ \frac{1-u}{C_N} & -\frac{a_0}{C_N} \end{bmatrix} \right| = a_0/C_N.$$

Suppose the current state is $J = (\theta_0, m = 0)$, and we propose to move to $K = (\theta_1^*, m^* = 1)$, then the acceptance probability for the move given by $\min(\alpha(J,K), 1)$, where

$$\alpha(J,K) = \frac{\pi(X|A,\theta_1^*,m^*=1)\pi(A|\theta_1^*,m^*=1)\pi(\theta_1^*|m^*=1)\pi(m^*=1)h'(u')m_{10}}{\pi(X|A,\theta_0,m=0)\pi(A|\theta_0,m=0)\pi(\theta_0|m=0)\pi(m=0)h(u)m_{01}}|J|, \tag{3.4.1}$$

where $\pi(\theta|m=i)$ is the joint prior distribution of θ specific to model i. We assign exponential prior distributions with rate λ to the transmission parameters, and note that the likelihood component $\pi(X|A,\theta,m)$ is not changed when moving from one model to another, since this does not involve the transmission parameters. This simplifies the acceptance ratio to

$$\alpha(J,K) = \frac{2}{3}\frac{\pi(A|\theta_1^*,m^*=1)\lambda^2\exp(-\lambda(a_0^*+a_1^*))}{\pi(A|\theta_0,m=0)\lambda\exp(-\lambda a_0)}\frac{a_0}{C_N}.$$

The reverse move is made by proposing a_0^* for model 0, given the current parameters a_0 and a_1 in model 1. This is proposed with the inverse of the transformation function g, which is

$$g^{-1}(a_0,a_1) = \left(a_0 + a_1C_N, \frac{a_0}{a_0+a_1C_N}\right) = (a_0^*, u').$$

The second component u' is ignored, as the dimension of the parameter space is reduced. The Jacobian is then $|J| = C_N/(a_0+a_1C_N) = C_N/a_0^*$.

Between models 1 and 2: This time, the candidate values $a_1^* = a_1(1+u)$ and $a_2^* = a_1(1-u)$ are proposed, where u is a random variable drawn from the Uniform(-1,1) distribution. The parameter a_0 is left unchanged. This transformation is equivalent to the diffeomorphism $g : [0,\infty)^2 \times [-1,1] \to [0,\infty)^3$, where

$$g(a_0,a_1,u) = (a_0, a_1(1+u), a_1(1-u)) = (a_0^*, a_1^*, a_2^*),$$

which has the Jacobian

$$|J| = \left| \det \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1+u & a_1 \\ 0 & 1-u & -a_1 \end{bmatrix} \right| = 2a_1.$$

Given states $J = (M_1, \theta_1)$, and $K = (M_2, \theta_2)$, then the acceptance probability for the move is

$$\begin{aligned}\alpha(J,K) &= \frac{\pi(X|A,\theta_2^*,m^*=2)\pi(A|\theta_2^*,m^*=2)\pi(\theta_2^*|m^*=2)\pi(m^*=1)h'(u')m_{21}}{\pi(X|A,\theta_1,m=1)\pi(A|\theta_1,m=1)\pi(\theta_1|m=1)\pi(m=1)h(u)m_{12}}|J| \\ &= \frac{3}{2}\frac{\pi(A|\theta_2^*,m^*=2)\lambda^3\exp(-\lambda(a_0^*+a_1^*+a_2^*))}{\pi(A|\theta_1,m=1)\lambda^2\exp(-\lambda a_0+a_1)}2a_1. \qquad (3.4.2)\end{aligned}$$

The reverse move is made by proposing a_0^*, a_1^* for model 1, given the current parameters a_0, a_1, a_2 in model 2. This is proposed with the inverse of the transformation function g, which is

$$g^{-1}(a_0,a_1,a_2) = \left(a_0, \frac{1}{2}(a_1+a_2), \frac{a_1-a_2}{a_1+a_2}\right) = (a_0^*, a_1^*, u').$$

The third component u' is ignored, as the dimension of the parameter space is reduced. The Jacobian is then $|J| = 1/(a_1+a_2) = 1/2a_1^*$.

If the algorithm proposes to stay in the current model, a standard within-model Metropolis-Hastings move is made. The Gibbs sampling step for p and z, as well as the data augmentation step are unaffected by the trans-dimensional moves suggested here. The efficiency of this algorithm depends on the transformation functions g, and the choice of C_N, which may be estimated prior to analysis, or the value at the current iteration could be used.

Convergence and mixing were monitored using trace plots, and running multiple chains from different starting points. Figures 3.2 and 3.3 show an examples of RJMCMC output for data simulated under model 0 and model 2 respectively. Within-model parameter estimates were compared to those derived from a regular within-model MCMC algorithm (such as those used in Chapter 2).

3.4.2 Factors affecting RJMCMC performance

A number of factors must be carefully considered before running the RJMCMC algorithm, to ensure efficiency and accurate results. In addition to issues common to RJMCMC analysis in all cases, certain properties relating to the transmission model can affect the performance of the algorithm. In this study, a series of simulation experiments were run to determine the impact of such factors on model selection. We simulated data according to each of the three MRSA transmission models, and then aimed to estimate the 'correct' model by assessing the posterior model probabilities. Unless otherwise stated, we generated 1000 patient admissions over 500 days, where $p = 0.07$, $z = 0.8$, and length of stay is drawn from a Poisson distribution with a mean

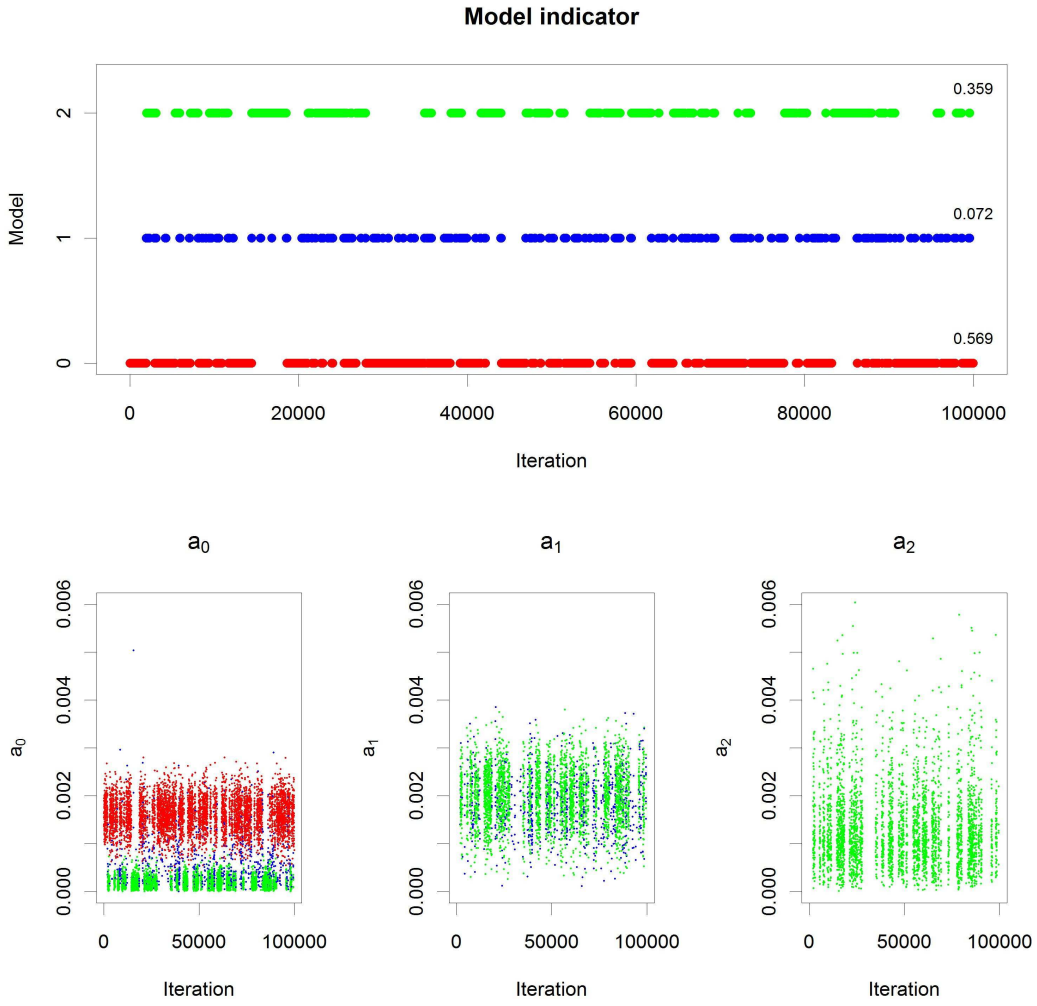

Figure 3.2: Trace plots for 100,000 iterations of the RJMCMC algorithm, having disregarded a burn-in sample of 20,000. Model indicator samples are shown in the upper plot, along with estimates for $\pi(m|x)$ for $m = 0, 1, 2$. The lower plots show samples for a_0, a_1 and a_2, with red indicating the current model $m = 0$, blue for $m = 1$ and green for $m = 2$. Data used here were simulated under model 0 with $a_0 = 0.002$. Exp(1) priors were assigned to the transmission parameters.

of 6 days. Under each scenario, we simulated and analysed several datasets. In the following presentation of results, we provide posterior model probabilities from one representative simulated dataset for each scenario, such that within-model parameter estimates were close to the values under which we simulated the data.

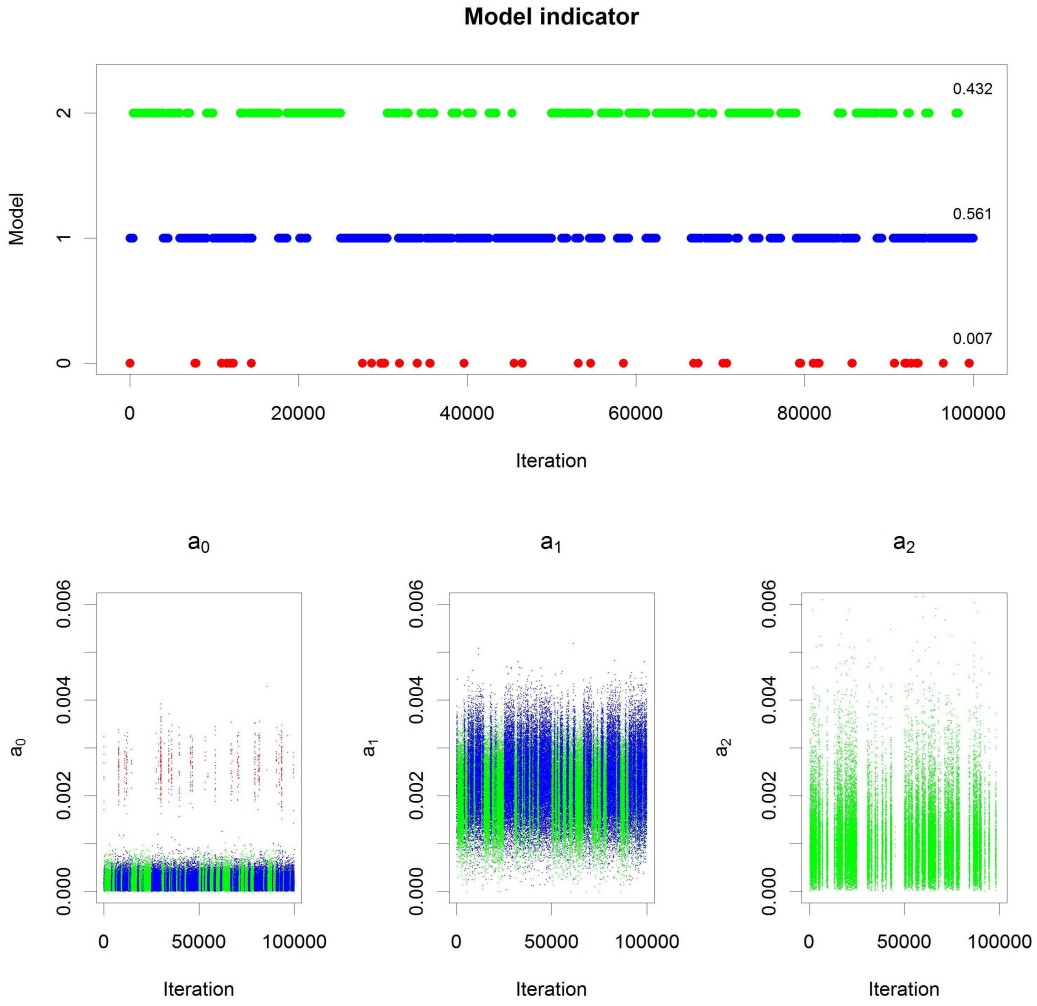

Figure 3.3: Trace plots for 100,000 iterations of the RJMCMC algorithm, having disregarded a burn-in sample of 20,000. Model indicator samples are shown in the upper plot, along with estimates for $\pi(m|x)$ for $m = 0, 1, 2$. The lower plots show samples for a_0, a_1 and a_2, with red indicating the current model $m = 0$, blue for $m = 1$ and green for $m = 2$. Data used here were simulated under model 2 with $a_0 = 0.0005$, $a_1 = 0.003$, $a_2 = 0.002$. Exp(1) priors were assigned to the transmission parameters.

3.4.2.1 Choice of prior distribution

The reversible jump MCMC process is typically very sensitive to the choice of prior distributions for the transmission parameters. Adding an additional parameter to a model will tend to adjust the prior odds in favour of the less complex model; this effect becomes greater as the prior distribution of this parameter becomes more diffuse. The influence of the prior distributions on the posterior model probability should be taken

into account. There are various measures that may be taken to avoid undue preference for simpler models. The use of informative priors can reduce this problem, and ideally a prior should be chosen which is informative enough to not affect the posterior model probability, but diffuse enough to have minimal influence on the posterior parameter estimates. Alternatively, prior matching may be used to minimise the discrepancy between the posteriors of models of different dimensions. Having chosen prior distributions for one model, we may attempt to choose prior distributions for a second model such that, *a priori*, the expected posteriors are as similar as possible.

We simulated data under the three alternative models described earlier. We generated datasets and ran the RJMCMC algorithm, varying the informativeness of the transmission parameter prior distributions, to determine the effect on the posterior model probability. The results are shown in table 3.1.

Prior effect on model posterior probability

m	Prior	a_0	a_1	a_2	$\pi(m=0\|x)$	$\pi(m=1\|x)$	$\pi(m=2\|x)$
0	Exp(10^{-3})	0.001	—	—	1	0	0
0	Exp(10^{-1})	0.001	—	—	0.98	0.01	0.01
0	Exp(1)	0.001	—	—	0.7	0.04	0.26
1	Exp(10^{-3})	0.0005	0.001	—	0.98	0.02	0
1	Exp(10^{-1})	0.0005	0.001	—	0.51	0.48	0.01
1	Exp(1)	0.0005	0.001	—	0.09	0.77	0.14
2	Exp(10^{-3})	0.0005	0.003	0.002	0.97	0.03	0
2	Exp(10^{-1})	0.0005	0.003	0.002	0.29	0.45	0.26
2	Exp(1)	0.0005	0.003	0.002	0.1	0.24	0.66

Table 3.1: RJMCMC results using data simulated under the models indicated in the first column. We generated 1000 patient admissions over 500 days, where $p = 0.07$ and $z = 0.8$. For each analysis, transmission parameters were assigned exponential prior distributions with rates 10^{-3}, 10^{-1} and 1 to determine the effect on the posterior model probability.

It is clear to see how simpler models are favoured while using diffuse priors. This suggests that one should be very cautious when using uninformative priors in a trans-dimensional setting. The impact of the informative prior on the posterior parameter estimates may be assessed by comparing them to within-model estimates with a diffuse prior. In this case, since the transmission rates are very small, Exp(1) is still fairly uninformative in the range of plausible values, meaning that posterior estimates are

only minimally affected. The posterior parameter estimates were compared to within-model estimates and very little difference was found, indicating that an Exp(1) prior distribution is not an unsuitable choice for these parameters.

3.4.2.2 Transformation functions

Transformation functions must be chosen carefully, in order to ensure that between-model parameter proposals are typically made to regions of relative high posterior probability density. This prevents the algorithm 'getting stuck' in one particular model, or group of models, due to implausible proposals.

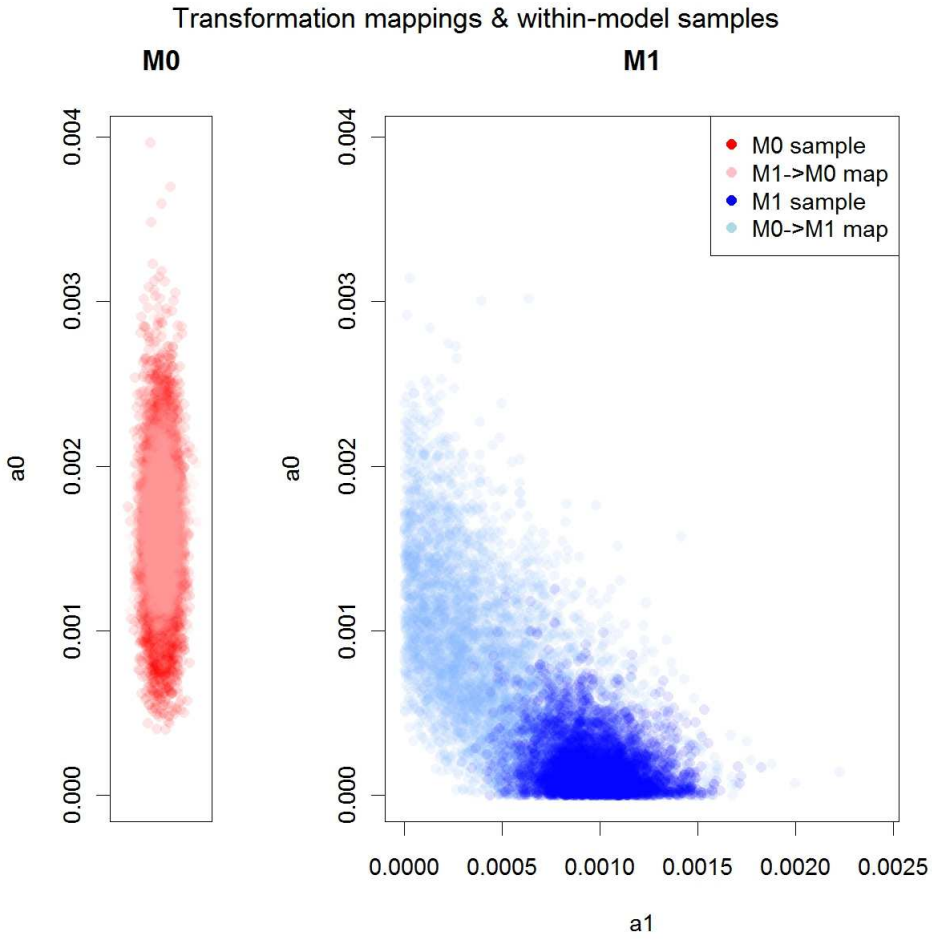

Figure 3.4: Within-model samples and transformation mappings between models 0 (red) and 1 (blue), for a simulated dataset. Darker points show samples from the within-model posterior distribution of the transmission parameters, while lighter points indicate the points proposed when a between-model jump is attempted.

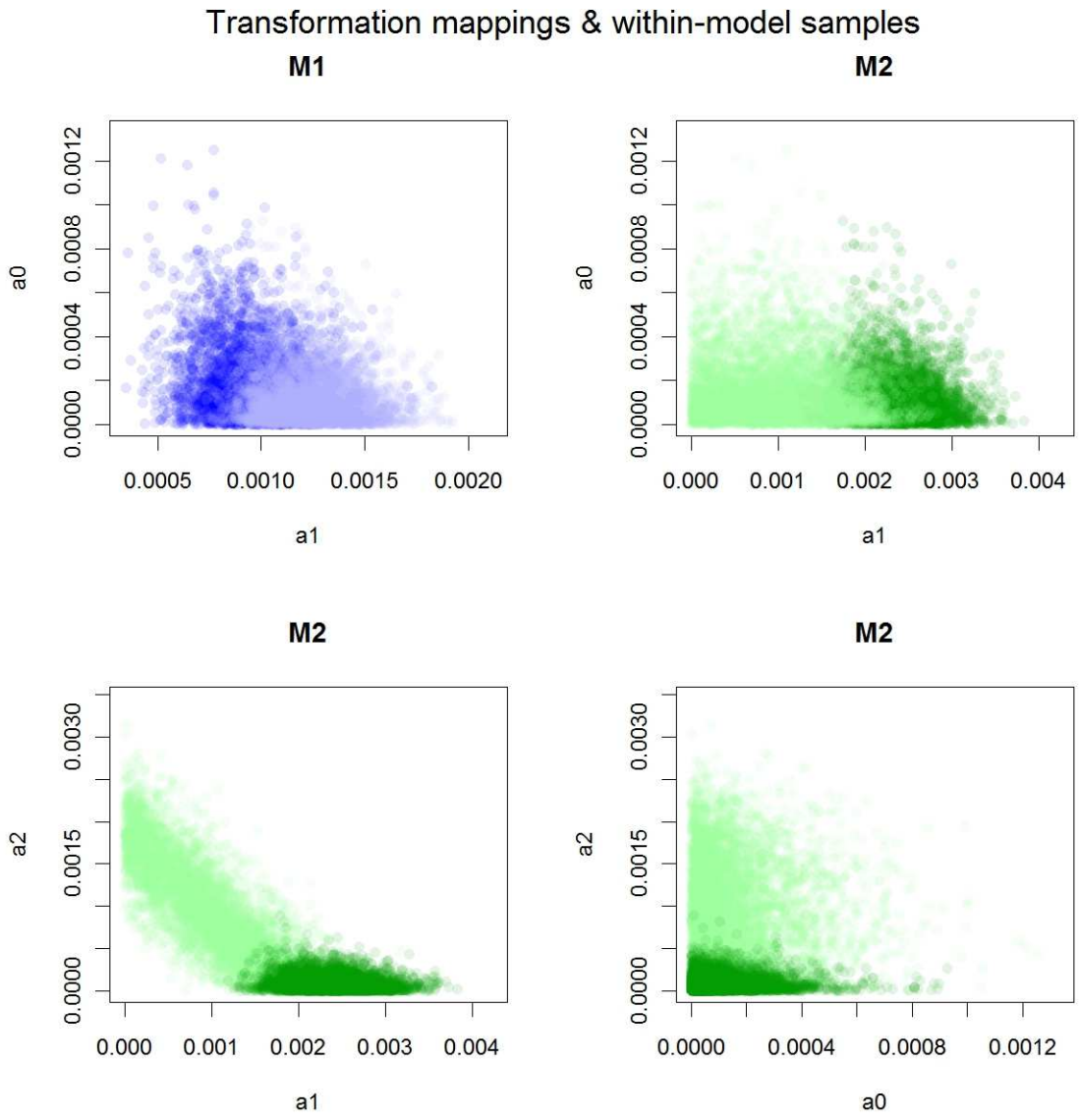

Figure 3.5: Within-model samples and transformation mappings between models 1 (blue) and 2 (green), for a simulated dataset. Darker points show samples from the within-model posterior distribution of the transmission parameters, while lighter points indicate the points proposed when a between-model jump is attempted.

The performance of the between-model parameter mappings proposed earlier was examined. Figures 3.4 and 3.5 show, for each model, within-model samples, and the proposed mappings to that model from a higher or lower dimension. We can see that proposals to a lower dimension model are very similar to accepted within-model points, which corresponds to a region of higher posterior probability. In contrast, proposals to a model of a higher dimension are frequently in regions associated with lower pos-

terior probability density, and are therefore very unlikely to be accepted. Proposing such redundant moves increases jump inefficiency, which causing the algorithm to get 'stuck' for large lengths of time. With such inefficiency, the algorithm must be run for a large number of iterations to avoid unduly favouring a particular model.

In order to combat this effect, the use of alternative transformations functions could be investigated, which more accurately reflect the relationship between the parameters in different models. The choice of transformation function in RJMCMC has been the topic of recent research [195–197]. The most successful (although perhaps not time-efficient) approach involves analysing the posterior parameter distribution for each model via pilot runs, and using this to inform the choice of proposal distribution. Clearly this becomes prohibitively time-consuming for a large model space. However, for a situation such as our MRSA transmission study with three candidate models, this could be performed relatively quickly — as noted in the previous chapter, within-model MCMC took over 6 hours to run.

An independence sampler can be constructed, where the proposed point does not depend on the current state of the Markov chain, but is drawn instead from a distribution determined by the pilot runs — for example, a Normal distribution could be fitted to the MCMC parameter samples, and a point proposed directly from this. Alternatively, a non-parametric approach could be taken, using kernel density estimation or a histogram fitted to the within-model parameter posterior, according to which across-model moves may be proposed.

3.4.2.3 Missing data

Further complications arise with the use of a missing data model, where the set of latent points A of unknown dimension is also being sampled. If the form of A is highly dependent on the model, then between-model jumps will be rare. Consider the set of colonisation times inferred during MCMC analysis of the MRSA transmission model. Figure 3.6 shows accepted values of A for each of the three candidate models, expressed as the count of acquisitions and importations. The areas of acceptance for each model appear highly distinct; the ratio of importations to acquisitions is higher for simpler models. These areas correspond closely to the region of high posterior probability for within-model analyses. In order to accept a point with a considerably lower likelihood than the current point, compensation must occur elsewhere in the acceptance ratio (eg. equations 3.4.1 and 3.4.2). As shown in figure 3.6, only a small proportion

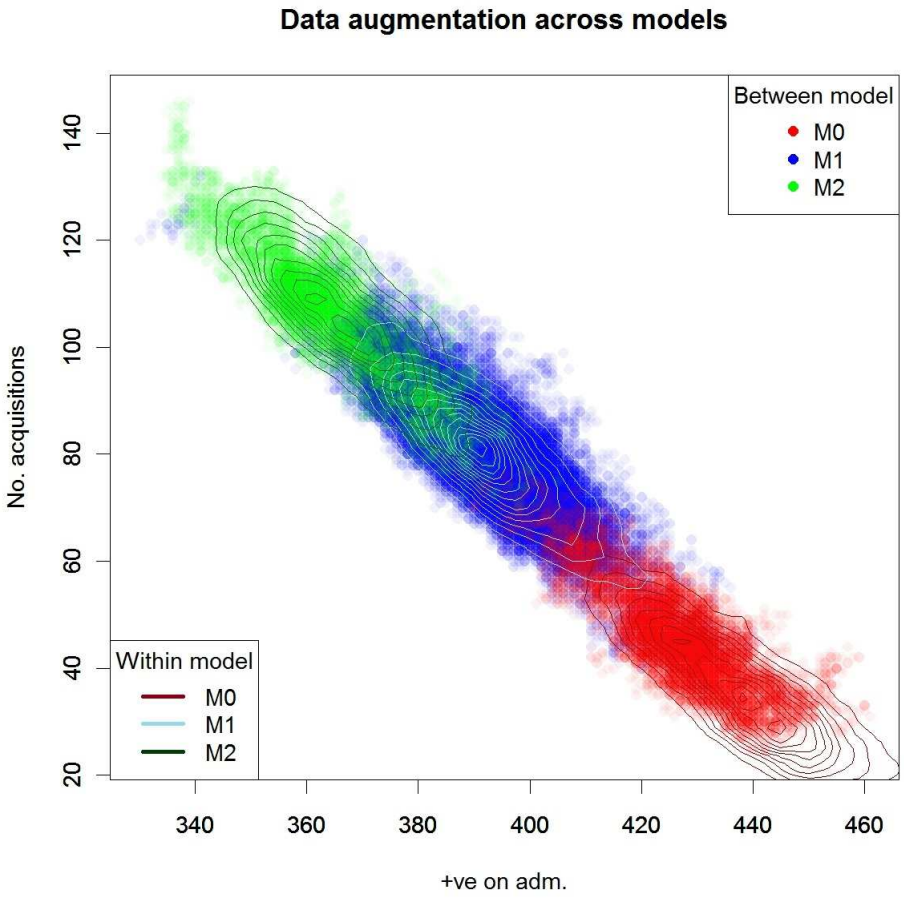

Figure 3.6: Accepted samples of the augmented data from the RJMCMC output are shown, where colours correspond to the associated model. Contour lines indicate the frequency of accepted samples from within-model analyses.

of proposed between-model jumps have any chance of acceptance, occurring when the value of A is in a region of high posterior probability for both the current and proposed model. While it may be possible to propose between-model moves for A which maintain reversibility, this is likely to be an extremely complex construction which may only marginally improve efficiency. One possibility is to allow the model jump probability $j(m,l)$ to depend on the 'distance' of the current state, A^*, from values of A associated with high posterior probability under model l. This should increase the number of between-model jumps by balancing out the discrepancy in likelihood values in the acceptance ratio, and the distance metric could be tuned to optimise the jump rate.

3.4.2.4 Rate of transmission

Since the differences between the models are based on the formulation of the transmission rate, a certain amount of colonisation events are necessary to inform our estimates. Datasets were simulated based upon the three candidate models, with varying levels of transmission, to find out in which scenarios the correct model might be chosen by the RJMCMC algorithm.

Effect of transmission rate on model posterior probability

Model, m	a_0	a_1	a_2	$\pi(m=0\|x)$	$\pi(m=1\|x)$	$\pi(m=2\|x)$
0	0.0005	—	—	0.99	0.01	0
0	0.001	—	—	0.7	0.04	0.26
0	0.005	—	—	0.59	0.04	0.37
0	0.01	—	—	0.68	0.02	0.3
1	0.0005	0.0005	—	0.99	0.01	0
1	0.0005	0.001	—	0.25	0.55	0.2
1	0.0005	0.002	—	0	0.47	0.53
1	0.0005	0.004	—	0	0.78	0.22
2	0.0005	0.001	0.0005	0.74	0.24	0.02
2	0.0005	0.002	0.001	0	0.8	0.2
2	0.0005	0.004	0.002	0	0.46	0.54

Table 3.2: RJMCMC results using data simulated based on models 0, 1 and 2, evaluated after 100000 iterations. 1000 patient admissions were simulated over 500 days with $p = 0.07$ and $z = 0.8$, and transmission parameters taking the values given in the table.

While a higher transmission rate seems to hinder the performance of the algorithm to correctly identify model 0, we found that high rates resulted in increased posterior probability of correctly choosing models 1 or 2. A very low transmission rate arising from any of the models is likely to convey very little information about transmission dynamics, and as such, a constant transmission rate is almost indistinguishable from anything more complex. Even with a high transmission rate, it was found that model 2 was not identified particularly well, with a similar posterior probability to model 1. The simulations under model 2 kept the ratio a_2/a_1 constant at 0.5, which might not be extreme enough to identify the more complex model under the simulation conditions. In order to determine if this is the case, a simulation experiment investigating

the impact of isolation effectiveness was also performed.

3.4.2.5 Isolation effectiveness

The identifiability of model 2, which separates the transmission effect of isolated and unisolated colonised patients, depends on the magnitude of this difference. In the case where $a_2 = a_1$, model 2 is identical to model 1, so poor selection might be expected when these parameters are very similar. To investigate this effect, datasets were simulated under model 2 with varying transmission parameters. Table 3.3 summarises the simulation results.

Effect of isolation effectiveness on model posterior probability

a_0	a_1	a_2	a_2/a_1	$\pi(m=0\|x)$	$\pi(m=1\|x)$	$\pi(m=2\|x)$
0.0005	0.002	0.002	1	0	0.9	0.1
0.0005	0.0025	0.002	0.8	0	0.47	0.53
0.0005	0.003	0.002	0.67	0	0.58	0.42
0.0005	0.002	0.001	0.5	0.02	0.87	0.11
0.0005	0.003	0.001	0.33	0.01	0.57	0.42
0.0005	0.004	0.001	0.25	0	0.01	0.99
0.0005	0.0025	0.0005	0.2	0	0.02	0.98

Table 3.3: RJMCMC results using data simulated based on model 2, evaluated after 100000 iterations. 1000 patient admissions were simulated over 500 days with $p = 0.07$ and $z = 0.8$, and transmission parameters taking the values given in the table.

It can be seen that the algorithm does not favour model 2 unless the transmission rate due to isolated patients is approximately four times less than that of unisolated patients. However, note that several interdependencies affect this performance, such as average length of stay and transmission intensity. It is also likely that increasing the study length will improve the correct identification of model 2 for smaller isolation effects.

3.4.2.6 Study length

Since the transmission rate of MRSA is typically low, and colonisation events are typically rare, data collected over a large time period is required to be able to investigate the underlying transmission model. Datasets were simulated for various parameter

values under each model, and for differing time spans. The posterior model probabilities were then analysed to determine how often the correct model was chosen; results are shown in table 3.4.

Effect of study length on model posterior probability

Study length	Model (m^*)	Proportion where $\pi(m^*\|x) > 0.5$	Proportion where $\pi(m^*\|x) > 0.75$
250	$m = 0$	0.6	0.55
	$m = 1$	0.4	0.3
	$m = 2$	0.25	0
500	$m = 0$	0.85	0.75
	$m = 1$	0.65	0.4
	$m = 2$	0.35	0.15
1000	$m = 0$	0.85	0.7
	$m = 1$	0.7	0.65
	$m = 2$	0.3	0.3
2500	$m = 0$	0.95	0.85
	$m = 1$	0.8	0.7
	$m = 2$	0.55	0.4

Table 3.4: Proportion of simulated datasets in which the posterior probability of the correct model, m^*, was greater than 0.5 and 0.75. 20 datasets were simulated with varying transmission rates for each level of study duration.

A general trend of increasing model identifiability associated with the increase in study length is be observed. Model 2 datasets were less frequently identified correctly, even for the longest study duration. Model 2 datasets simulated with similar values for a_1 and a_2 tended to have a higher posterior probability of model 1, as discussed earlier.

3.4.3 Analysis of GST data

The MRSA surveillance data collected at Guy's and St. Thomas' hospital, London (GST), described previously in section 2.3, is now considered.

The RJMCMC algorithm was run for 100000 iterations using various priors for the transmission parameters. The results are summarised in table 3.5. Transmission parameters were assigned an exponential prior distribution with rates 1 and 10^{-3}, and the strong effect of the uninformative prior distribution on the posterior model proba-

GST posterior model probabilities

	Exp(1)			Exp(10^{-3})		
Ward	$m = 0$	$m = 1$	$m = 2$	$m = 0$	$m = 1$	$m = 2$
1	0.02	0.41	**0.57**	0.25	**0.7**	0.05
2	0	**0.75**	0.25	**0.52**	0.33	0.15
3	**0.98**	0.02	0	**0.99**	0.01	0
4	0.05	0.42	**0.53**	**0.56**	0.43	0.01
5	**0.66**	0.33	0.01	**0.81**	0.18	0.01
6	0.03	0.23	**0.74**	**0.69**	0.12	0.19
7	**0.89**	0.07	0.04	**0.98**	0.01	0
8	0.24	**0.75**	0	**0.8**	0.19	0.01
9	**0.95**	0.05	0	**1**	0	0
10	**0.9**	0.04	0.06	**0.94**	0.06	0

Table 3.5: The posterior model probabilities under different prior assumptions for the transmission parameters. For each ward, the highest posterior probability is shown in bold.

bility is observed again. The posterior probability of model 0 was found to be greater than 0.5 in five of the ten study wards, even under the informative prior. On the whole, these were the wards with the lowest acquisition rates, which, as discussed in the previous section, may lead to favouring the basic, constant rate model.

There was not a particularly high support for model 2 in general, the model incorporating isolation effect. The simulation studies showed that model 2 is only detectable when isolation effect is clear — a reduction of approximately 75% was required to observe a high posterior probability for this model with a study length of 500 days. This corresponds to the higher posterior probability of model 2 seen for wards 1 and 4, which were also estimated to have a high isolation effectiveness in the previous chapter.

The within-model analyses indicated that wards 2 and 6, the elderly care wards, were similar in terms of demographics, transmission rate, and typical length of stay. However, the reversible jump model selection results for these two wards look quite different. This suggests that the underlying model for each dataset is not the same, indicating model 1 is more likely for ward 2, and model 2 more likely for ward 6. Figures 3.7(a) and 3.7(b) shows the estimated isolated and unisolated colonised populations for wards 2 and 6 respectively. These populations are typically similar in ward 2, but

greater differences may be observed in ward 6. If these populations are the same, then one cannot differentiate between models 1 and 2. This may partly explain the difference in posterior model probability between these two datasets.

The GST study lasted for 16 months. Via simulation studies in the previous section, it was found that model selection was not particularly successful for studies of a similar length, especially where transmission was very low. A short study length and low transmission rates may cause more complex models to appear indistinguishable from simpler formulations. One should therefore be cautious in the interpretation of these results. RJMCMC is likely to perform better on larger datasets than these, or in settings with a greater rate of transmission.

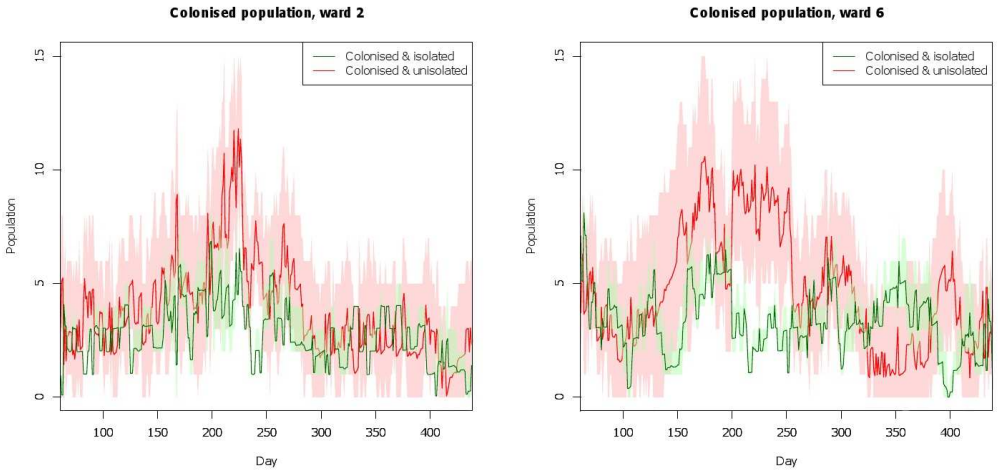

Figure 3.7: Colonised populations for wards 2 and 6 (elderly care), showing number of isolated (green) and unisolated (red) patients, with shaded 95% credible intervals.

3.5 Deviance information criterion

We now consider the use of the DIC to select a model which fits the data best, in a parsimonious manner. In our setting, we have missing data, and use of the DIC as specified by Spiegehalter et al. [136] is not possible. Instead, we adopt the DIC_6 measure described in section 3.2.4.2, and henceforth refer to this measure simply as the DIC.

For this procedure, we initially run a within-model data-augmented MCMC algorithm for each model for 100,000 iterations, in order to derive posterior mean values for p

and z, and median values for transmission parameters. Transmission parameters were assigned exponential prior distributions with rate 10^{-3}. Having derived posterior estimates $\tilde{\theta}$, we ran the algorithm a second time for 100,000 iterations, with $\theta = \tilde{\theta}$, allowing only the augmented data to vary. This allowed us to derive the DIC for a given model, as described earlier in section 3.2.4.2. As with the RJMCMC procedure, there are a number of factors which will affect the outcome of model selection via DIC. We conducted simulation studies to determine the factors which affect the performance of the DIC in identifying the 'correct' model. We denote DIC(m) to be the value of the DIC for model m. We simulated data in the same manner as during the investigation of RJMCMC performance. In the presentation of our results, we provide the DIC values calculated from one representative simulated dataset for each scenario.

3.5.1 Factors affecting DIC outcomes

3.5.1.1 Choice of prior

The first simulation study was undertaken to analyse the performance of the DIC measure under different prior assumptions for the transmission parameters. Results are displayed in table 3.6.

Prior effect on DIC outcome

Model, m	Prior	a_0	a_1	a_2	DIC(0)	DIC(1)	DIC(2)
0	Exp(1)	0.001	—	—	**867.2**	902.1	922.0
0	Exp(0.1)	0.001	—	—	**866.0**	900.3	921.6
0	Exp(0.001)	0.001	—	—	**857.5**	895.8	915.8
1	Exp(1)	0.0005	0.001	—	**1376.4**	1410.6	1400.2
1	Exp(0.1)	0.0005	0.001	—	**1379.6**	1408.8	1395.1
1	Exp(0.001)	0.0005	0.001	—	**1344.8**	1403.8	1390.1
2	Exp(1)	0.0005	0.002	0.003	3341.4	3170.3	**3165.4**
2	Exp(0.1)	0.0005	0.002	0.003	3338.2	3207.8	**3164.7**
2	Exp(0.001)	0.0005	0.002	0.003	3205.6	3167.8	**3162.1**

Table 3.6: DIC results using data simulated based on models 0, 1 and 2. For each analysis, transmission parameters were assigned prior distributions to varying degrees of diffuseness to determine the effect on the posterior model probability.

It was found that increasing prior uncertainty of the transmission parameters reduced

the value of the DIC under all models. While the value of the DIC is affected by the choice of prior distribution, the conclusion remains the same — the simplest model is supported under data generated from models 0 and 1, while model 2 is correctly favoured in the dataset simulated under model 2. While the DIC clearly fails to correctly identify model 1 in all cases, it is observed that the selection of model 2 is unaffected by the prior. The DIC measure avoids Lindley's paradox, which is an advantage over calculating posterior model probabilities.

3.5.1.2 Rate of transmission

The prevalence of carriers in the ward at any given time affects the expected number of transmission events for all but the simplest model. This in turn drives the amount of information available to estimate the transmission dynamics. A further simulation study was conducted to determine the effect of the transmission rate on model selection via the DIC measure.

Effect of transmission rate on DIC

Model, m	a_0	a_1	a_2	DIC(0)	DIC(1)	DIC(2)
0	0.0005	—	—	**806.1**	849.8	861.8
0	0.001	—	—	**870.1**	921.2	931.6
0	0.005	—	—	**1609.0**	1643.4	1638.7
0	0.01	—	—	**2392.6**	2471.3	2459.8
1	0.0005	0.0005	—	**891.7**	906.6	909.7
1	0.0005	0.001	—	**1032.8**	1093.6	1100.0
1	0.0005	0.002	—	2477.5	2395.4	**2384.6**
1	0.0005	0.004	—	4328.7	**4290.9**	4291.8
2	0.0005	0.001	0.0005	**1178.2**	1194.5	1197.6
2	0.0005	0.002	0.001	2130.4	2131.8	**2128.8**
2	0.0005	0.004	0.002	4146.6	4074.5	**4067.3**

Table 3.7: DIC results using data simulated based on models 0, 1 and 2. Data were simulated using various transmission parameters, which are given in the table.

Table 3.7 shows that model selection is generally more successful under higher transmission rates. While model 1 was identified correctly in a high transmission scenario, the DIC values under models 1 and 2 was very similar. While it is difficult to interpret

the difference between DIC values, this seems to suggest that both models provide a similar fit to data generated by model 1.

3.5.1.3 Isolation effectiveness

As mentioned previously, the degree to which models 1 and 2 differ is given by the scale of isolation effectiveness, or the relative magnitude of a_1 to a_2.

Effect of isolation effectiveness on DIC

a_0	a_1	a_2	a_2/a_1	DIC(0)	DIC(1)	DIC(2)
0.0005	0.002	0.002	1	2549.0	2517.4	**2517.1**
0.0005	0.0025	0.002	0.8	2774.3	2691.7	**2681.7**
0.0005	0.003	0.002	0.67	2696.8	2577.7	**2548.1**
0.0005	0.002	0.001	0.5	2200.1	**2166.4**	2179.7
0.0005	0.003	0.001	0.33	3120.4	3030.2	**3029.1**
0.0005	0.004	0.001	0.25	2538.8	2451.5	**2427.5**
0.0005	0.0025	0.0005	0.2	2163.0	2137.6	**2108.7**

Table 3.8: DIC results using data simulated based on model 2, under various values of a_1 and a_2, as indicated in the table.

Table 3.8 seems to suggest that model selection via DIC is successful over a range of different transmission values for a_1 and a_2, as in most cases, model 2 is correctly selected. However, it seems that datasets generated under model 1 also result in a DIC selection of model 2. The previous simulation study demonstrated that selection of model 1 was poor, and the results seen in this study seem to exhibit the same effect.

3.5.1.4 Study length

Since MRSA transmission rates are typically very low, and study populations are small, we require data collected over a long period to observe a sufficient number of acquisitions with which to investigate the underlying model. We ran a simulation study to determine how much data we might have to collect in order to correctly determine the model with the DIC.

Table 3.9 illustrates the need for a long period of data collection in order for correct models to be identified by DIC, under plausible transmission rates. The proportion of datasets generated under model 1 which were correctly identified (that is, had the

Effect of study length on DIC outcome

Study length	Model (m)	Proportion where DIC(m) is lowest
250	0	1
	1	0
	2	0.15
500	0	0.95
	1	0.05
	2	0.25
1000	0	1
	1	0.1
	2	0.45
2500	0	0.95
	1	0.1
	2	0.75

Table 3.9: Proportion of simulated datasets in which the correct model, m, was selected (took the lowest DIC value). 20 datasets were simulated with varying transmission rates for each level of study duration.

lowest DIC value under model 1) was particularly low, even for a study length of 2500 days.

3.5.2 Analysis of GST data

Having run several simulation studies on the model selection performance of DIC_6, it has been observed that unless transmission rates are very high (table 3.7), or that study length is very long (table 3.9, it is not possible to make any distinction from the most basic, constant rate model with this approach. Since the dataset collected from GST spans approximately 500 days, it might be concluded that unless transmission rates are particularly high, it is very likely that the simple model will be selected by DIC_6. The DIC was calculated with two MCMC runs of 100,000 iterations, the second run fixing θ at the posterior means. Transmission parameters were assigned exponential prior distributions with mean 1000. Results are summarised in table 3.10.

It was found that model 0 was the clear selection for all wards. This is almost certainly due to the relatively short study length and low transmission rate of the observed data,

DIC values for GST data			
Ward	DIC ($m = 0$)	DIC ($m = 1$)	DIC ($m = 2$)
1	**750.6**	821.4	830.2
2	**861.5**	934.8	945.9
3	**1907.2**	1975.2	2004.5
4	**1862.3**	1928.8	1931.7
5	**477.2**	824.3	760.0
6	**982.3**	1053.0	1083.9
7	**925.4**	1032.1	1049.7
8	**700.0**	750.8	771.3
9	**351.4**	387.7	397.9
10	**590.3**	621.5	622.4

Table 3.10: DIC values for GST data for each of the candidate models.

which were below the levels we found to be necessary via simulation studies to be able to differentiate between models using the DIC.

3.6 Conclusion

We have shown that Bayesian model choice for hospital transmission models requires careful use of model selection methods, which are sensitive to many different factors. It seems clear that the power to detect a difference between models is strongly dependent on the number of transmission events inferred from the data, and therefore, the study length and transmission rate. Both RJMCMC and DIC approaches failed to identify models when the dataset was collected over a shorter periods of time (250-500 days); the performance of DIC was particularly poor. As pointed out in the original paper, DIC is not applicable in all cases, and the epidemic models used here generate a highly complex likelihood over parameters and missing data [136]. The measure p_D is shown to be a good approximation of effective number of parameters when there is a well-behaved likelihood function, but may fail in a more complex setting. Indeed, some of the simulation studies generated low and even negative values for p_D, indicating that using the DIC may not be entirely appropriate. This indicates that the expected likelihood with θ fixed at the posterior mean $\bar{\theta}$ is lower than the expected likelihood over the augmented data and the parameters.

A major issue in RJMCMC is the selection of prior distributions for the model parameters, as the outcome is highly dependent on this choice. Lindley's paradox means that the more uninformative the parameter prior distributions, the greater the support given to simpler models. This leads to a trade-off between within and between-model estimation — an uninformative prior is desired to estimate the transmission parameters, but will lead to posterior model probabilities weighted in favour of simple models. When using more informative priors, the sensitivity of the posterior parameter estimates to this should be considered.

It would be of interest to explore techniques to mitigate the effect of uninformative prior distributions. For instance, in our setting, we may attempt to specify prior distributions for the transmission parameters such that the induced prior distribution of the overall transmission rate is approximately similar across each model. This might be achieved by matching the prior expectation and variance across models. Consider the *a priori* transmission rates for models 0 and 1:

$$\begin{aligned} E(q_0(t)) &:= E(q_1(t)) \\ E(a_0^{(0)}) &= E(a_0^{(1)}) + n_C E(a_1^{(1)}), \end{aligned}$$

where we indicate the model number as a superscript for clarity, and n_C denotes the average number of colonised patients. Similarly,

$$\mathrm{Var}(a_0^{(0)}) := \mathrm{Var}(a_0^{(1)}) + n_C^2 \mathrm{Var}(a_1^{(1)}).$$

By specifying 'equivalent' prior distributions for parameters in each model, the algorithm is less likely to reject moves to models of higher dimension.

An advantage of the RJMCMC procedure is that both posterior model probabilities and within-model parameters are estimated in the same process. While the posterior probability of a model is easily understood and compared, the DIC value provides little information beyond one model being better than another to some degree. In a Bayesian framework, it is natural to consider all models as possible, to varying degrees of plausibility. It could be argued that DIC is considerably 'less Bayesian' for this reason. Furthermore, the DIC uses point estimates ($\tilde{\theta}$) in its calculation, which violates the Bayesian paradigm.

The posterior model probabilities may be used in a Bayesian model averaging procedure. For instance, in the setting we have been considering in this chapter, we may easily derive a weighted average of the transmission rate, or number of acquisitions, using the output from the RJMCMC algorithm. This allows us to provide estimates for

quantities of interest using a range of possible models, rather than having to specify one particular model.

In terms of ease of calculation, the DIC has some advantages. A RJMCMC analysis is not a black-box procedure, and must be carefully tailored to the individual situation, in order to monitor and optimise between-model jumps, and to assess the sensitivity to prior information. As discussed earlier, one can avoid the issue of constructing efficient transdimensional jumps by running within-model analyses to evaluate the regions of high posterior probability. With a large model space, this becomes impractical, and the DIC may become an attractive alternative.

Spiegelhalter et al. made it clear that the DIC is not universally applicable, and recommend its use in cases with approximately normal likelihoods, where the posterior mean provides a good estimate, pointing out that non-log-concave likelihoods could provide negative values for p_D. This paper generated controversy upon publication [136, discussion]. Alongside much praise for the development of such a model selection criterion, criticism was aimed at the "arbitrary assumptions and approximations" (Brooks), the non-Bayesian nature of using a point estimate ($\tilde{\theta}$) in a supposedly Bayesian measure, and the lack of applicability of the process to more general modelling situations. Additionally, since the DIC has no scale, it is not possible to estimate how much better one model is than another, and one can do little more than select the model with the lowest DIC value. However, in providing a model choice statistic which is, in many cases, easier and quicker to calculate than calibrating and performing a RJMCMC analysis, the DIC is a useful model comparison tool, particularly when the set of models to compare is large.

While it was found that model selection was fairly poor using either method on small datasets, or those where transmission was typically low, RJMCMC proved to be the more successful method, as long as fairly informative prior distributions are specified (Exp(1), for example).

It is worth noting that models with lower posterior probability or higher DIC values should not necessarily be cast aside, as they may still be used to provide useful or informative estimates. It is, however, of great importance to determine that the chosen model provides a good fit to the data. This is pertinent to model 2 described in this chapter, as it provides estimates of clinical interest (isolation effectiveness), but seems not to be selected for many of the wards via RJMCMC, or at all via DIC. We found in the previous chapter that this model provided a good fit to the data, using the posterior

predictive distribution. We must also bear in mind the limitations to both model choice approaches, which have a large impact on the choice of model for small to medium-sized datasets, such as those we have analysed.

Having systematically considered the performance of both RJMCMC and the DIC in stochastic epidemic model selection, it is clear that the amount of transmission data available is key to being able to differentiate between candidate models. Low transmission rates mean that infection events are rare and often unclustered, and as such, it is not possible to differentiate this activity from a constant rate Poisson process, without very long studies. We found that with a longer average length of stay, there was a greater chance of selecting the correct model.

CHAPTER 4

Inference of transmission using whole genome sequence data

4.1 Introduction

While many studies have investigated the dynamics of MRSA transmission in hospitals, estimating transmission rates and the effectiveness of various infection control measures, uncertainty about the true routes of transmission remains. Using conventional screening methods, it is difficult to predict the source of a given MRSA-positive individual's colonisation. The ability to estimate transmission routes may increase understanding of how a pathogen spreads in hospital wards, and could potentially reduce the uncertainty of parameter estimates.

In section 4.2 we consider existing methods to analyse genetic data, and studies which utilise such data to provide an insight into epidemic behaviour and transmission dynamics. Section 4.3 introduces a dataset collected from a hospital in Thailand, in which MRSA isolates were taken from colonised patients and sequenced over a period of three months. We describe some of the issues with analysing these data, in particular, modelling the transmission process at a genetic level, and the within-host dynamics.

Following this, two methods are introduced to analyse this dataset, each with a different aim. The first method is described in section 4.4, in which we aim to estimate any differences in transmissibility by grouping MRSA isolates into genetically similar groups, and considering each to have a different transmission rate. The second approach is described in section 4.5, in which the aim is to estimate a transmission network, while simultaneously estimating transmission parameters and other quantities

of interest. We describe a simulation study to assess the performance of this method, and consider methods to measure the accuracy of an estimated network. In section 4.6, the results of both of our analyses are presented. Finally, section 4.7 is a discussion of the findings, and how our approach compares with previous studies, as well as potential future extensions to this work.

4.2 Background

4.2.1 Transmission networks and phylogenetic trees

A transmission network is a graph representing the spread of a pathogen on an individual level. It comprises nodes, representing infected individuals, and directed edges, representing transmission events. Edges may additionally be associated with a transmission time. A transmission network may be composed of multiple unconnected subnetworks, each representing independent outbreaks of a disease. Each connected transmission subnetwork (or 'transmission chain') has an origin, representing the original introduction of the pathogen into a population. In many epidemic studies where the entire infected population is considered, it is sensible to regard the network as fully connected (that is, only one origin exists). However, when considering dynamics within a subpopulation, such as a hospital or ICU, it may be realistic to expect multiple introductions of the pathogen from external sources. In this chapter, a transmission network is regarded as one or more subnetworks, allowing for multiple introductions, as is common in a healthcare setting.

A phylogenetic tree, or genealogy, describes the relatedness of a set of organisms, and represents evolutionary divergence over time. The tips of the tree represent observed individuals, and internal nodes represent divergences (sometimes called bifurcation or speciation events), which are typically unobserved. These represent hypothetical ancestors of the observed organisms. Nodes directly linked by one internal node are said to share an ancestor, and are genetically similar relative to the rest of the population. Genetic distance between a node and its ancestor may be represented by the length of the edge.

There are strong similarities between the transmission network and the phylogenetic tree, or genealogy. After a transmission event between individuals A and B, the pathogen found present on B can be thought of as a descendent of the pathogen colonizing A. While 'descendant' is rather poorly defined when discussing colonies of several

thousand organisms with a short generational time, the transmission network can be thought of as a simplification of the genealogy, where the ancestors of individual B's colony were found colonizing individual A at the time of transmission.

Figure 4.1 shows a phylogenetic tree derived from MRSA isolates collected in an ICU (data described later in section 4.3.1). The tree indicates clusters of genetically similar isolates, which are additionally given similar colours. Horizontal branch lengths have been scaled to indicate genetic distance. While phylogenetic trees might be used to rule out transmission routes between genetically distant isolates, more information would be required to estimate a transmission network. The tree in figure 4.1 was generated by optimising a likelihood function of the sequence data, described in further detail in the next section. The optimisation was performed using the `ape` package for `R 2.12.1` [198].

Whether interest lies in the genealogy or the transmission network, the exact structure is almost always unobserved, and the true structure must be inferred, based on a set of observations. Over the past decade, attempts have been made to combine the analysis of evolutionary dynamics with transmission dynamics, using genetic data.

4.2.2 Phylogenetic tree reconstruction

Some of the earliest work on phylogenetic reconstruction was done in the 1960s, prior to the availability of genetic sequence data. Edwards and Cavalli-Sforza described the method of minimum evolution, in which the most plausible tree is given as that which represents the 'least total evolution', or genetic change [199]. This is somewhat loosely defined, but can be applied to phenotypic traits — roughly speaking, an organism's ancestors are likely to share similar characteristics, and can be classified as such.

Genetic data provides a quantitative basis to demonstrate similarity. Using basic distance measures, there are various methods to reconstruct the phylogenetic tree. The simplest approaches are clustering algorithms, which find and group genetically similar pairs, such as the unweighted pair group method with arithmetic mean (UPGMA) and neighbour-joining method [200]. These methods iteratively find and combine the most genetically similar pair of nodes/individuals, treat this pair as one node, and then recalculate the distances from this new node to all others before repeating the process, each time reducing dimensionality by one.

A maximum likelihood approach, incorporating methods to search tree topologies, was

described by Felsenstein in 1981 [201]. Suppose a tree is composed of a set of nodes $X = \{X_1, \ldots, X_n\}$. The edge function $e(A, B)$ returns 1 if A and B are linked by an edge on the tree, and 0 if not. Each node is associated with a time, which might represent an observation or generation time. For each $A, B \in X$, τ_{AB} is the time elapsed between the time associated with A and B. Each node is represented by a genome of length N, and the ith nucleotide of the sequence A is given at $A^{(i)}$. Assuming that mutations occur independently and at random, and that the sequences represented by all internal and external nodes are known, then the model is defined by the joint probability mass

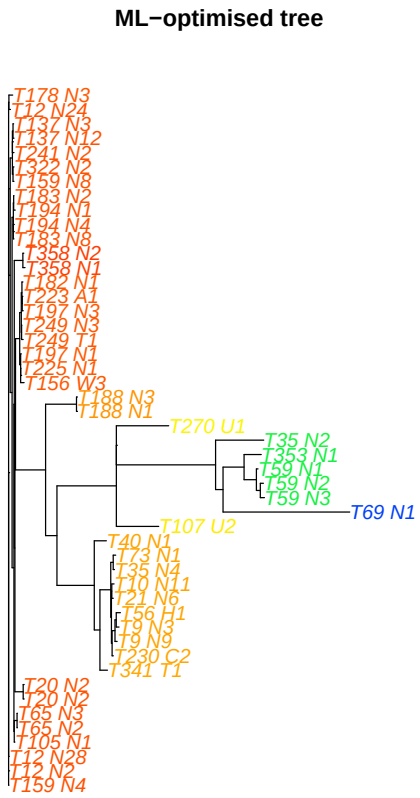

Figure 4.1: An example of a phylogenetic tree, derived with a maximum-likelihood approach. Horizontal branch length is proportional to genetic distance; genetically similar observations have a similar label colour. The data used to generate this tree were collected from a paediatric ICU in Thailand, described in section 4.3.1.

function

$$f(X|\theta) = \prod_{\substack{A,B\in X \\ e(A,B)=1}} \prod_{l=1}^{N} \Big[\mathbf{1}_{A^{(l)}\neq B^{(l)}} P(\text{mutation from } A^{(l)} \text{ to } B^{(l)} \text{ after time } \tau_{AB}|\theta) + \mathbf{1}_{A^{(l)}=B^{(l)}} P(\text{no mutation at } l \text{ after time } \tau_{AB}|\theta) \Big].$$

However, since the sequences realised at the internal nodes are typically unobserved, we must sum over all possible states at these points. Felsenstein proposed the 'pruning algorithm' to calculate this sum, and proposed a topology searching method to estimate an optimal tree [201].

4.2.3 Coalescent theory

Coalescent theory has been used extensively in the past decade to infer past population dynamics from current observations. This method first requires a genealogy to be determined, by one of the many methods of tree reconstruction. By doing this, the reproduction times, or branching events, may be estimated. A genealogy describes the generation of organisms from an origin individual at time 0, to n contemporary individuals at some time T. It is typically assumed that the n sequences are a random sample taken from a larger population. A total of $n-1$ bifurcation events occur in the interval $[0, T]$, increasing the sample population by 1 each time. Coalescent theory considers this process in reverse; we have a set of n contemporary sequences at time 0, and suppose $n-1$ coalescent events occur until there is one remaining current ancestor at time T. Let $A(t)$ be the number of lineages (number of ancestors in the sample population) present at time t. The sequence $\{A(t)\}$ may be modelled as a stochastic process, decreasing from n to 1 [202]. According to a simple coalescent model, the probability of a coalescent event occurring in a small time interval $[t, t+\delta t]$ is given as

$$P(A(t+\delta t) = i-1 | A(t) = i) = \binom{i}{2} \frac{1}{N(t)} \delta t + o(\delta t),$$

where $N(t)$ is a relative size function, which may be assumed to be proportional to the population size — as such it is commonly referred to as the effective population size [202]. This may be defined as any demographic model reflecting the population over time; for example, a constant rate, or exponential growth. In an epidemic setting, $N(t)$ may be thought of as the effective number of infectives at time t, since this population is driving the rate of infections (coalescent events). The parameters of the demographic model may be evaluated via the likelihood of observing a set of coalescent times, given the underlying population model.

Pybus et al. introduced a simple epidemic model structure to estimate the past population dynamics of the Hepatitis C virus (HCV) [203]. This method does not allow transmission parameters to be estimated; primary interest in this study was in the past population dynamics of the pathogen. However, the authors demonstrate how the basic reproduction number R_0 may be estimated with this approach.

Several papers since have used and extended this methodology to describe the past population size and evolutionary dynamics of pathogens [202, 204, 205].

Rasmussen et al. described a method to analyse epidemiological data and phylogenetic structure simultaneously [206]. The authors construct a likelihood function with two components, one for each data type, which are assumed to be independent. Epidemiological data are assumed to be in the form of imperfectly-observed infection counts over time, while sequence data are used to construct a genealogy, from which one obtains coalescent times. The authors describe a likelihood contribution from each of these data sources. The coalescent rate with i ancestors present at time t is given as

$$\frac{\binom{i}{2}}{\binom{I(t)}{2}}\beta(t)\frac{S(t)}{N}I(t),$$

where $\beta(t)$ is a transmission function, and $I(t)$ is the imputed number of infective individuals at time t. The authors use a particle MCMC approach to sample approximately from the posterior distribution, thus avoiding computation of the intractable likelihood. The analysis requires the phylogenetic tree to be estimated pre-analysis, incorporating no uncertainty. In this study, interest lies primarily in the population dynamics inferred from both data sources.

4.2.4 Reconstructing transmission networks

Cottam et al. provided a method to link the analysis of genetic and epidemiological data, with the aim of providing a most likely transmission network [151]. The authors define a gamma distribution describing the incubation period based on estimates from existing literature, and a probability mass function for the time of infection, based on observational data. These are used to calculate the probability that any given individual was infectious at a particular time. The probability of a specific transmission route is then dependent on the relative infectiousness of potential sources at the time of infection. Sequence data are used to construct a set of plausible transmission trees,

for which the likelihood may be calculated according to the probability functions described above. Those with higher likelihoods may be selected as plausible transmission networks. This approach is applied to an outbreak of foot-and-mouth disease in the UK, with data collected from 20 farms. In this approach, sequence data are only used to generate a set of most plausible trees, narrowing down the vast set of all trees. The likelihood of the epidemiological data is then calculated for each of these trees. Clearly this method becomes much more computationally demanding for larger datasets. Their findings are also dependent on the choice of distribution for incubation periods.

Jombart et al. described methods to investigate emerging epidemics using sequenced isolates collected at various times during the outbreak [152]. This approach optimises a weighted transmission network constrained by possible transmission times. A network is chosen such that the number of single nucleotide polymorphisms (SNPs) between nodes (individuals) is minimal. In the case of multiple edges representing equal genetic distance, the edges are weighted to be time-dependent, and the number of SNPs is assumed to follow a Poisson distribution with mean $\mu L \Delta_t$, where μ is the prespecified mutation rate of the organism, L is the length of the genome, and Δ_t is the time between samples from the two connected nodes. This method allows the analysis of sequence data collected during an epidemic, but does not take into account the collection of multiple sequences from one individual over time. Furthermore, transmission between person i and j is assumed to occur strictly after the sampled sequence from i; thus, if the sample time of i is before that of j, j cannot be the source of infection for i. The algorithm deterministically returns the most plausible network, and does not indicate any degree of uncertainty associated with this structure.

Ypma et al. described a Bayesian approach to transmission network reconstruction [65]. Data were collected from Dutch poultry farms to study an outbreak of avian influenza, including sequence data for selected genes. Between-farm transmission was modelled according to farm infectiousness (which was constant until culling took place, at which time it reduced exponentially), spatial distance and genetic similarity. These components were assumed to be independent. Since transmission dynamics are unobserved, the authors describe an MCMC algorithm to sample over transmission routes (infection trees), and model parameters. The resulting transmission network is assumed to be fully connected.

4.2.5 Aims

Previous studies have operated under assumptions which are restrictive and sometimes unrealistic, and are incompatible with the analysis of MRSA transmission in hospital wards. We aimed to avoid these issues, and provide the following:

1. A flexible model describing within-host and between-host genetic behaviour, under which we may simulate data in order to investigate the performance of the model under various scenarios.
2. A model which incorporates multiple sequences collected over time from each positive individual.
3. A model which allows for frequent introductions of the pathogen to the population from newly admitted patients, which results in multiple unconnected transmission subnetworks.

In this chapter, we aimed to use both whole genome sequence (WGS) data and epidemiological data to address two different questions about MRSA transmission in hospital wards. Firstly, we investigated heterogeneity in MRSA transmission rates according to genetic type, using a clustering method. Secondly, we aimed to recover the unobserved transmission network in the hospital ward. We present these methods in sections 4.4 and 4.5 respectively, but first, we introduce the data and notation which are used in this chapter.

4.3 Analysis of genetic data in a hospital setting

4.3.1 Data

We used MRSA surveillance data[1] collected from a hospital in North-East Thailand. A study was conducted in two intensive care units (ICUs), specialising in paediatrics (ICU 1) and surgery (ICU 2). ICU 1 recorded 170 admissions (169 unique patients), while ICU 2 had 114 admissions (98 unique patients) over a period of three months, during which time individuals were regularly screened at various body sites for carriage of MRSA. A total of 1640 patient days were recorded across both wards. In total,

[1]These data have not previously been published. Many thanks to S. Peacock, M. Holden, E. Nickerson, M. Hongsuwan, J. Parkhill and others involved with data collection and processing for the provision of this dataset.

140 positive isolates were collected, of which 83 were sequenced — 51 different patients had at least one MRSA isolate sequenced. One patient in particular (T126) had 19 sequenced isolates taken at regular intervals. A summary of the patients participating in the study is given in table 4.1, and details of colonised patients are shown in figures 4.2 and 4.3. Across all 83 sequenced isolates, polymorphisms were detected at 2591 different loci.

Summary of Thai ICU data

	ICU 1	ICU 2
Ward type	Paediatric	Surgery
Number of patient episodes	170	114
Number of unique patients	169	98
Number of episodes with ≥ 1 positive swab	20	29
Total number of positive swabs collected	51	89
Total number of positive swabs sequenced	43	40
Mean length of stay (days)	4.6	7.8

Table 4.1: Summary of patients admitted during the ICUs during the three month study.

4.3.2 Notation

Suppose there are a total of n patient admissions to a hospital, with admission days $t_1^a, \ldots, t_n^a$, and discharge days $t_1^d, \ldots, t_n^d$. Multiple readmissions for one individual were considered to be independent. Each patient j was screened for MRSA carriage ν_j times, $\nu_j \geq 0$, on days $t_{j,1}^x, \ldots, t_{j,\nu_j}^x$, with (positive or negative) results $x_{j,1}, \ldots, x_{j,\nu_j}$. In addition, MRSA positive patients have ρ_j ($0 \leq \rho_j \leq \nu_j$) isolates taken for sequencing on days $t_{j,1}^y, \ldots, t_{j,\rho_j}^y$, with whole genome sequences $y_{j,1}, \ldots, y_{j,\rho_j}$. Sequence times are a subset of positive screen times, so that $0 \leq \rho_j \leq \nu_j$. A patient j has an (unobserved) colonisation time given by t_j^c, which takes a value of ∞ if the patient remains uncolonised. If patient j is colonised on admission, then $t_j^c = t_j^a$, and we set importation marker $\phi_j = 1$; otherwise, $\phi_j = 0$.

Genetic similarity is measured by the number of SNPs between two isolates. The number of SNPs between isolates a and b is denoted $\psi(a, b)$.

A full table of the notation used in this chapter is provided on page 179.

4.4 Assessing heterogeneity in transmissibility of MRSA strains

In the first modelling method, we aimed to estimate transmission rates from genetically distinct groups of MRSA isolates. We supposed that G groups of genetically similar strains of MRSA are observed, each of which is associated with a certain level of transmissibility. We supposed that each group of patients colonised by a particular

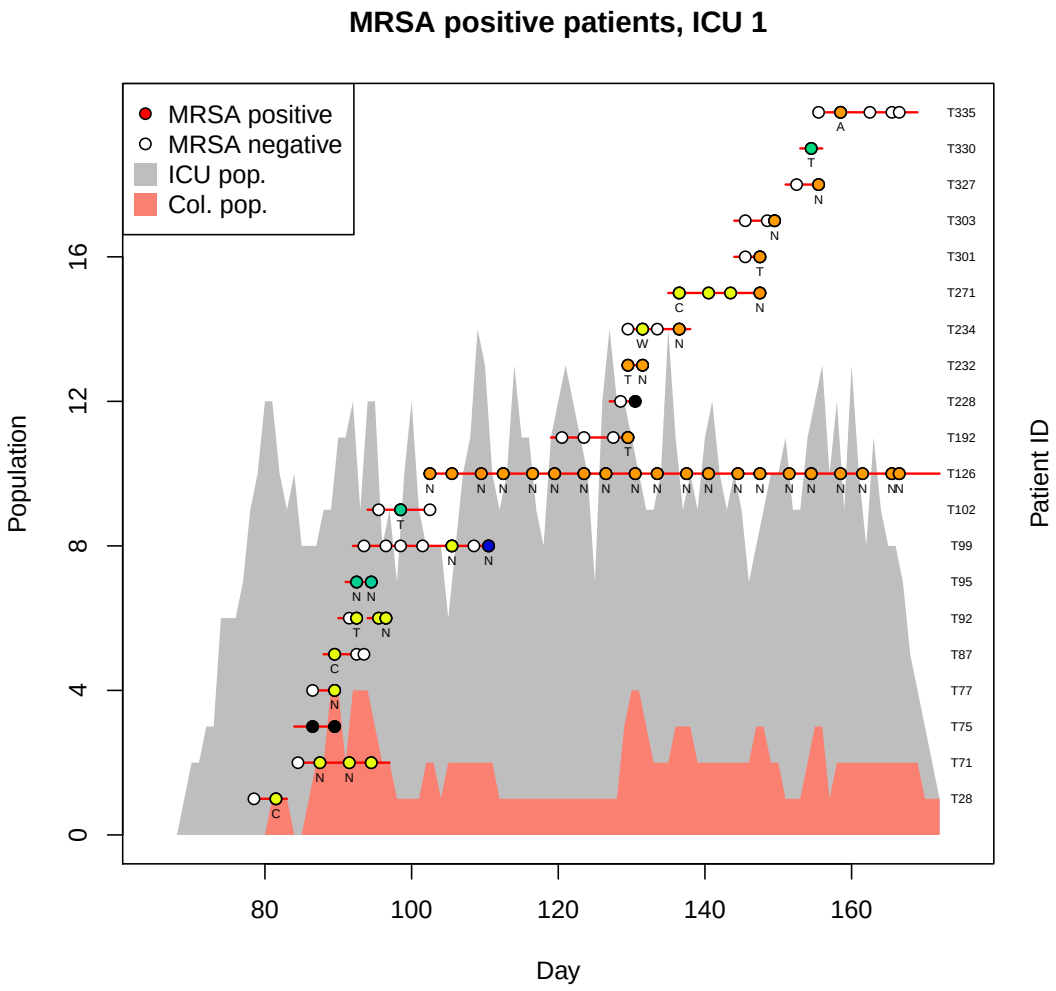

Figure 4.2: Colonised (pink) and total (grey) population over time in ICU 1. Also plotted are the patient episodes for individuals with at least one MRSA positive swab at some point. White points are MRSA negative swabs, and coloured points are positives. The colours indicates genetic distance; similar colours represent genetically similar isolates. Untyped MRSA isolates are assigned the same colour as the last typed isolate; those isolates of unknown type are coloured black. The letter 'N' indicates a sequenced nasal swab, 'A' axilla, 'T' throat, 'C' trachea, 'W' wound.

MRSA type exerts transmission pressure on each susceptible patient independently. Susceptible patients are supposed to be homogeneous in terms of susceptibility. We define the transmission rate for a given susceptible individual as

$$q_G(t) = \sum_{i=1}^{G} a_i C_i(t),$$

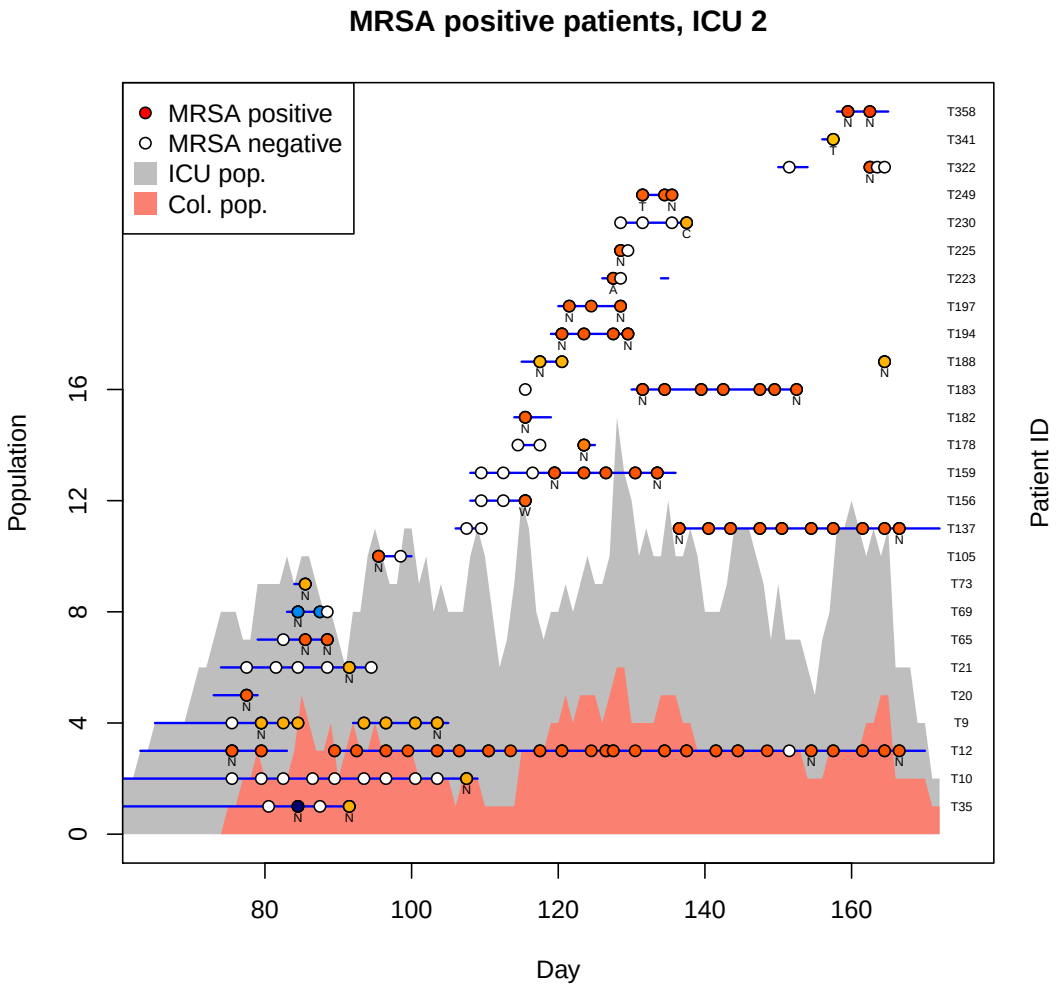

Figure 4.3: Colonised (pink) and total (grey) population over time in ICU 2. Also plotted are the patient episodes for individuals with at least one MRSA positive swab at some point. White points are MRSA negative swabs, and coloured points are positives. The colours indicates genetic distance; similar colours represent genetically similar isolates. Untyped MRSA isolates are assigned the same colour as the last typed isolate; those isolates of unknown type are coloured black. The letter 'N' indicates a sequenced nasal swab, 'A' axilla, 'T' throat, 'C' trachea, 'W' wound.

where $C_i(t)$ is the number of patients present in the ward on day t, colonised by a strain of MRSA in group i. Let $C(t) = \sum_j C_j(t)$ be the total number of colonised patients present on day t. We worked in discrete time, and assumed that patients colonised on day t contribute to the colonised population (of their MRSA type) from day $t+1$ until their discharge. Patients positive on admission join the colonised population on the day of admission. The probability of a susceptible becoming colonised by an MRSA carrier in group i on day t is then

$$P(\text{colonised on day } t) \cdot P(\text{first effective contact on day } t \text{ with group } i)$$
$$(1 - \exp(-q_G(t)) \cdot \frac{a_i C_i(t)}{q_G(t)}.$$

We assumed that a patient colonised with MRSA of type i must have acquired this from another type-i positive individual in the hospital, or imported this strain from outside the ward. It was also assumed that once colonised, a patient remained colonised by the same type of MRSA. This excludes the possibility of loss of carriage, large-scale within-host mutations, concurrent carriage of multiple genetically diverse strains, or reinfection by another type. It was found that few patients exhibited evidence for carriage of highly diverse types; this is investigated in more detail later, in section 4.5.2.

A similar model to those used in previous chapters is adopted — the framework is described in section 2.4.4. Let p be the probability of being colonised on admission, and z be the sensitivity of the swab test. Let g_j be the MRSA group to which patient j belongs, and $g = \{g_1, \ldots, g_n\}$. If patient j is not colonised, set $g_j = 0$. The parameter space is augmented with the set of unobserved data $A = \{\phi, t^c, g\}$, consisting of the unobserved colonisation times t^c, admission statuses ϕ ($\phi_j = 1$ if patient j positive on admission, otherwise 0) and MRSA groupings g, where $g_j \in \{1, \ldots, G\}$ for each colonised patient. By doing so, the full likelihood $\pi(X, A|\theta)$ becomes tractable, and a data-augmented MCMC process is used to sample from the posterior distribution of θ and A.

4.4.1 Data augmentation

At each iteration of the MCMC algorithm, we choose uniformly at random to add, delete or move a colonisation event, and propose new values ϕ^*, t^{c*}, g^*. We define $A^* = \{\phi^*, t^{c*}, g^*\}$ to be the new dataset proposed by this sampling step. We define the proposal ratio $q_{A,A^*} = P(A^* \to A)/P(A \to A^*)$, which is the ratio of probabilities

of making the reverse move and the proposed move. Patients inferred to be MRSA positive without observed sequences are assigned an MRSA type at random. Let v_s be the number of patients never screened positive, v_q be the number of patients who carry MRSA at some point during their episode (either observed, or added by the algorithm), and v_a be the number of patients for whom a colonisation time has been added by the algorithm. We note that v_s is fixed, while v_q and v_a are updated as the algorithm progresses. With equal probability, one of the following moves is selected:

- **Change colonisation time.** Select at random one of the v_q patients with a colonisation time. With probability w, propose the selected patient j was positive on admission, otherwise sample a colonisation time t_j^{c*} from $\{t_j^a, \ldots, l_j\}$, where l_j is the last potential day of colonisation (the earliest from day of discharge and day of first positive screen). If $C_{g_j}(t_j^{c*}) = 0$, no move is made. For this move we have

$$q_{A,A^*} = \begin{cases} 1 & \text{acquisition - acquisition} \\ \frac{1-w}{w(l_j - t_j^a + 1)} & \text{acquisition - importation} \\ \frac{w(l_j - t_j^a + 1)}{(1-w)} & \text{importation - acquisition} \\ 1 & \text{importation - importation} \end{cases}$$

  MRSA groups remain unaffected by this move.

- **Add colonisation.** Select at random one of the $v_s - v_a$ patients who are currently assumed to be negative. If $v_s - v_a = 0$, then no move is made. With probability w, add an importation, otherwise add an acquisition. If an importation is proposed for the selected patient j, set $\phi_j^* = 1$, $t_j^c = t_j^a$, and draw the MRSA group uniformly at random from $\{1, \ldots, G\}$. If an acquisition is proposed, then draw a colonisation time t_j^{c*} from $\{t_j^a, \ldots, t_j^d\}$. Select an MRSA group g_j^* at random with probability $C_{g_j^*}(t_j^{c*})/C(t_j^{c*})$. If no colonised patients are present, then no move is made.

- **Remove colonisation.** Choose at random one of the v_a patients who have had a colonisation time added by the data augmentation process. If $v_a = 0$, no move is made.

Having established the augmented data move mechanisms, the probability ratios q_{A,A^*} for adding or removing colonisation times may be given as follows:

	Importation	Acquisition
Add	$\frac{(v_s-v_a)G}{w(v_a+1)}$	$\frac{(v_s-v_a)(t_j^d-t_j^a+1)C(t_j^{c*})}{(1-w)(v_a+1)C_{g_j^*}(t_j^{c*})}$
Remove	$\frac{v_a w}{(v_s-v_a+1)G}$	$\frac{v_a(1-w)C_{g_j}(t_j^c)}{(t_j^d-t_j^a+1)(v_s-v_a+1)C(t_j^c)}$

Having sampled a candidate colonisation time/source, the proposed augmented dataset A^* is accepted with probability

$$\min\left(1, \frac{\pi(X|A^*,\theta)\pi(A^*|\theta)}{\pi(X|A,\theta)\pi(A|\theta)} q_{A,A^*}\right).$$

4.4.2 Assigning groups

Groups are determined by assigning membership to isolates in such a way as to minimise the total within-cluster genetic distance. Suppose we observe a total of n_s sequences, $y_1, \ldots, y_{n_s}$. Let A_G be a set of clusters, $K_1 \ldots, K_G$, where each cluster K_i contains one or more isolates. We define the total within-group distance as

$$V(A_G) = \sum_{i=1}^{G} \sum_{0<j<k\leq n} \mathbf{1}_{y_j,y_k\in K_i} \psi(y_j, y_k),$$

where $\psi(y_j, y_k)$ is the number of SNPs between isolates y_j and y_k. Then the optimal clustering, A_G^*, is the grouping which minimises within-group distance:

$$A_G^* = \underset{K_1,\ldots,K_G}{\arg\min}\, V(A_G).$$

This is a variation of the k-means clustering approach [104]. In order to achieve the optimal grouping, the following algorithm is performed:

k-means clustering algorithm for DNA sequences

1. Of the $n_s = \sum_i \rho_i$ available sequences $y_1, \ldots, y_{n_s}$, choose randomly G sequences to be the representative ('mean') sequences for clusters $1, \ldots, G$. These are denoted $x_1, \ldots, x_G$.

2. Assign each sequence y_j a cluster membership g_j by choosing its closest cluster representative,

$$g_j = \underset{i=1,\ldots,G}{\arg\min}\ \psi(y_j, x_i).$$

3. For each cluster i, reselect a representative strain x_i which minimises the total distance to all other sequences in the group;

$$x_i = \underset{x}{\arg\min} \sum_{x_k : g_k = i} \psi(x, x_k).$$

   This may be achieved by minimising the distance of each nucleotide to all members of the group independently.

4. Repeat steps 2 and 3 until convergence.

This algorithm assigns each of the n_s DNA sequences a group label, such that there are G clusters with minimal within-cluster variation.

It is worth noting that this approach does not specifically require whole genome sequence data, and can be performed with lower resolution genetic data, such as that generated by MLST or *spa* typing.

4.4.3 Number of clusters

It is not obvious what the most appropriate choice of G should be, as this depends on the amount of available data, the amount of total variation existing across clusters, and any genetic factors which may contribute to transmissibility. Ideally isolates should be parsimoniously partitioned into groups which have low variation. It is clear that $V(A^*_{G+1}) \leq V(A^*_G)$ for $G = 1, \ldots, n-1$. The minimised distance for between one and

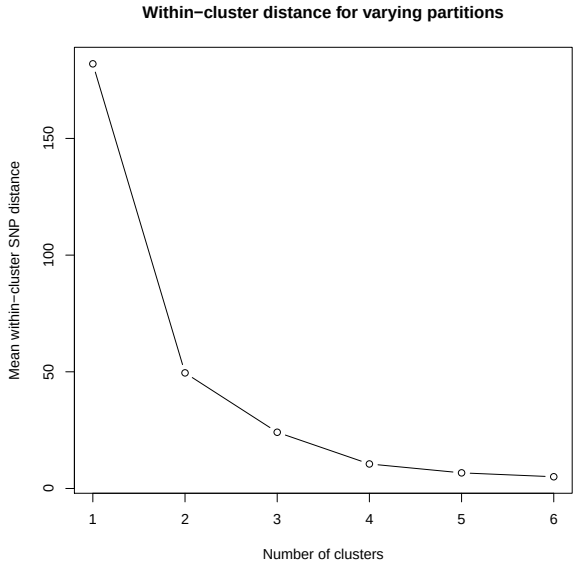

Figure 4.4: The average distance (number of SNPs) of each isolate from another member of its cluster, where the number of clusters varies between 1 and 6.

six clusters using the Thai data is shown in figure 4.4. While increasing the number of clusters reduces the variation within groups, this also increases the number of transmission parameters to be estimated. A small number of isolates informing a parameter estimate will typically result in a large degree of uncertainty of the estimate.

While any choice of G is to some degree arbitrary, we chose to explore different models with G ranging from 1 to 4. Alternatively, one could specify a threshold level of genetic similarity (in terms of SNPs) determining group membership.

4.5 Estimation of transmission networks

In the second modelling method, we aimed to estimate the transmission network within each of the ICUs, while estimating key transmission parameters. We assumed the transmission dynamics within the hospital are represented by a collection of one or more trees, where nodes represent positive patients, and edges are transmission routes. Each tree has a distinct origin, which is an importation from outside the hospital, since we assume that background transmission is not possible. A transmission network is completely defined in this setting by $T = \{\phi, t^c, s\}$, where ϕ is the set of admission statuses (tree origins), t^c the set of colonisation times (branching times), and s the transmission sources (directed edges) for all patients.

In this section, we firstly introduce some extra notation, which we require to discuss

transmission networks based on sequence data. We discuss genetic diversity and how we might attempt to model the diversity within-host and between individuals in 4.5.2. We describe our modelling framework, and define the two models which we use for our analysis in 4.5.3. We then describe the data augmentation process which we use to sample over the unobserved transmission times and routes. We describe how we can simulate data under our models, and how we can use this to assess the performance of our methods in a variety of scenarios in 4.5.6.

4.5.1 Notation for genetic data

In addition to the definitions in section 4.3.2, we require some additional notation for our network reconstruction approach.

For a positive patient j, a subset of ρ_j positive MRSA isolates are sequenced, at times $t^y_{j,1}, \ldots, t^y_{j,\rho_j}$, resulting in a set of whole genome sequences $y_j = \{y_{j,1}, \ldots, y_{j,\rho_j}\}$ (an empty set if $\rho_j = 0$). In total, a total of $\sum_i \rho_i = n_s$ whole genome sequences are collected. For convenience, all sequences are ranked according to date collected. Sequences collected on the same day are ordered arbitrarily. We label the sequences $1, \ldots, n_s$, and define $r(k)$ to be the patient ID associated with the kth sequence.

We denote the genetic distance between two sequences A and B in terms of the number of SNPs, by $\psi(A, B)$. Let Ψ be the symmetric $n_s \times n_s$ matrix of pairwise genetic distances, such that $\Psi_{i,j} = \psi(i, j)$. Clearly $\Psi_{i,i} = 0$ for all i, and we have $n_s(n_s - 1)/2$ unique pairwise distances.

Each colonised patient j has a source s_j equal to the ID of the patient from whom the colonisation was acquired, or $s_j = 0$ if the MRSA colonisation has not been acquired from another patient in the ward (that is, an imported strain).

Let $t(i, j)$ be the function which describes the length of the path between nodes i and j in an unweighted network, ignoring edge direction. Since individuals can only have one source of colonisation, transmission networks are acyclic and there exists one unique path between any pair of nodes. We have that $t(i, j) = t(j, i)$ for all nodes i and j. Furthermore, if i and j belong to the same tree, then $0 \leq t(i, j) < \infty$, with $t(i, j) = 0$ if and only if $i = j$. If i and j do not belong to the same tree, then we set $t(i, j) = \infty$. In our setting, $t(i, j)$ represents the number of transmission events between patients i and j. If i is colonised by j, then $t(i, j) = 1$. An example is shown in figure 4.5.

A full table of the notation used in this chapter is provided on page 179.

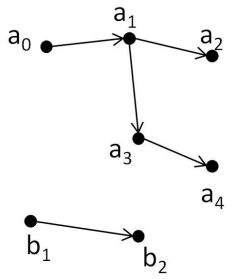

Figure 4.5: Transmission network, consisting of two subtrees, arrows denoting transmission between patients (dots). Patient a_0 and a_4 are three transmission events apart ($t(a_0, a_4) = 3$), as are patients a_2 and a_4. Expected genetic diversity grows as the number of transmission events in a chain increases. Patients b_1 and a_1 are unrelated, so $t(a_1, b_1) = \infty$.

4.5.2 Modelling genetic variation

In order to successfully integrate genetic data into the analysis of pathogen transmission dynamics, it is important to consider the processes occurring at the molecular level, as well as the individual and population level. WGS data provide an insight into the genetic similarity between isolates at a much higher resolution to earlier typing methods, such as *spa* typing and MLST. As such, it has only recently become possible to study *in vivo* bacterial microevolution, and the molecular level processes are still incompletely understood [38].

4.5.2.1 Within-host variation

Two isolates taken from the same individual at times t_1 and t_2 may differ for numerous possible reasons:

1. At the moment of colonisation, multiple genetic types have been transferred to the individual, and these coexist within the same inoculum. At each swab time, one isolate is sequenced, which is a sample from this diverse population.

2. The individual is originally colonised with a single genetic type of *S. aureus*, and mutations occur within this population, resulting in either a stable mixed population of genetic types (so that repeated samples are effectively sampling from the same population), or a population made up of genetic types, whose number and proportions change over time (a different population may be sampled at

each time point). Populations may change due to genetic drift, or as a result of fitness differences. These effects may be indistinguishable over short periods of time. If a patient is positive on admission, they may have been colonised for a long time prior to observation; as such, may potentially have a greater degree of within-host diversity than a newly-colonised individual.

3. An individual may be colonised a second time between time t_1 and t_2, resulting in a different bacterial population profile. This new type may replace or coexist with the original colony.

4. Isolates may be taken from different body sites. Independent, genetically distinct colonies may exist due to separate colonisation events.

5. Contamination or sequencing errors may lead to two genetically identical strains appearing to differ.

The mutation rate for MRSA has been estimated to be 3.3 (2.5, 4) $\times 10^{-6}$ per site per year [66], which corresponds to around 9-10 mutations per year with a genome length of 2.8Mb [79]. Although it was once assumed that colonisation was typically due to a single type, carriage of multiple strains of *S. aureus* has been detected using non-WGS typing methods [207, 208], and variation is much more likely to be detected with higher resolution sequence data. Young et al. conducted a study into evolutionary dynamics of *S. aureus*, considering in particular the microevolution occurring within-host during the progression from carriage to infection [38]. WGS data collected from three participants revealed a degree of within-host genetic diversity, which would not be detected with conventional genetic typing methods. A total of 30 SNPs were observed from the 68 sequenced isolates taken from the participant acquiring *S. aureus* infection, while 42 and 39 SNPs were recorded from the asymptomatic carriers. Isolates were taken over several months, and the rate of evolution was estimated to be 2.7 $\times 10^{-6}$ mutations per site per year. This study indicated that large-scale mutations were unnecessary to cause infection, and that individuals may be colonised by a several genetically similar strains.

Figure 4.6 shows the observed genetic distance over time for within-host pairs of isolates for the Thai datasets. For each individual i with two or more sequences, the genetic distance between the jth and kth sequence ($\psi(y_{i,j}, y_{i,k})$) is plotted against the time between the collection of these sequences ($|t_j^y - t_k^y|$). There is no evidence for increasing within-host diversity over the short period covered in this study, meaning that it

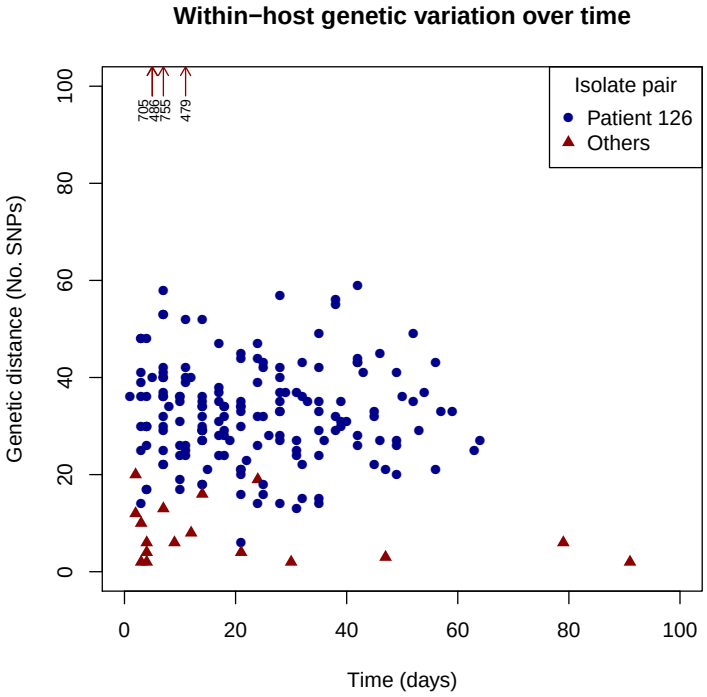

Figure 4.6: For all patients with multiple isolates, the genetic distance of each isolate pair is plotted against the time between samples. 19 isolates were available for patient T126 (resulting in 171 within-host swab pairs). All other patients had fewer than 4 sequenced isolates each. Four outlying points are represented by arrows subtitled with the corresponding SNP count.

is unlikely that the observed within-host genetic diversity is due to genetic drift alone. Patient T126 had 19 sequenced isolates, and each pairwise genetic distance is indicated in this plot. There appears to be a slightly greater diversity within this individual than others, although the data available for other patients are limited. Note the outlying points indicated in figure 4.6. These represent four individuals with highly dissimilar isolates 5-10 days after an initial sample. Two of these individuals (patients T234 and T271, see fig. 4.2) had isolates taken from different body sites, and the lack of genetic similarity may be attributed to independent colonisations. The other two are pairs taken from the same body site, but have a magnitude far greater than the large majority of pairs ($>$ 500 SNPs compared to 10-20). This may point towards a subsequent colonisation event between sample times, or possibly a large-scale genetic change in the pre-existing colony. Bacterial recombination events result in large genetic distances (in terms of SNPs) emerging in a short period of time. Any advantage in fitness of this new type may lead to increasing prevalence, and possibly replacement of an older type. Estimating the degree to which recombination affects genetic variation in bacteria, and

detecting evidence for recombination events, is challenging [209]. Castillo-Ramírez recently estimated that slightly over half of the variation found in MRSA ST239 caused by mutation is due to recombination [210]. Croucher et al. estimated that a polymorphism was around seven times more likely to have been introduced through recombination than point mutation in *Streptococcus pneumoniae* [211]. Methods for identifying recombination have developed alongside the increasing resolution of genetic data. Hein [212], and more recently Marttinen et al. [213] and Didelot et al. [214] described methods for identifying recombination events; a similar approach was presented by Lawson et al. [215] with the aim of inferring population structure.

Without multiple sequences for contemporaneous isolates, it is difficult, or perhaps impossible, to estimate the genetic diversity across a bacterial colony, and indeed, how this might change over time. Diversity due to sampling and due to mutation events appear to be indistinguishable with these data, so we made no attempt to separate these effects while pursuing the primary goal of estimating transmission networks.

4.5.2.2 Transmission chain diversity

In a typical transmission event, a small number of the bacteria cells hosted by a colonised patient are transferred to another, often via the hands of a healthcare worker, possibly attached to a flake of skin, or present in bodily fluid. The initial colony is likely to be small. Upon colonisation, the bacteria enter a lag phase, as they adjust to a new environment, before growing at an exponential rate [168]. Genetic diversity can established in the new host during this growth period, resulting in a new colony which has a different composition to that of the source. A transmission event is a population bottleneck, in which only a small sample of a population survive and propagate (in the newly-colonised individual). Genetic diversity is typically reduced as a result of a population bottleneck [216]. However, we observed a similar degree of within-host diversity amongst all patients in the Thai dataset, which could indicate that diversity is acquired at the time of transmission, or that diversity is established shortly after transmission (see figure 4.7).

In their study on avian flu in Dutch farms, Ypma et al. supposed that nucleotide substitutions occur at different frequencies [65]. The number of transitions (d_{ts}) and transversions (d_{tv}) (see discussion in section 1.3.1) between an infected farm and its source were recorded. In addition, the observance of a genetic deletion is denoted with the indicator function d_{del} (equal to 1 if deletion occurs, 0 otherwise). They provide a likelihood

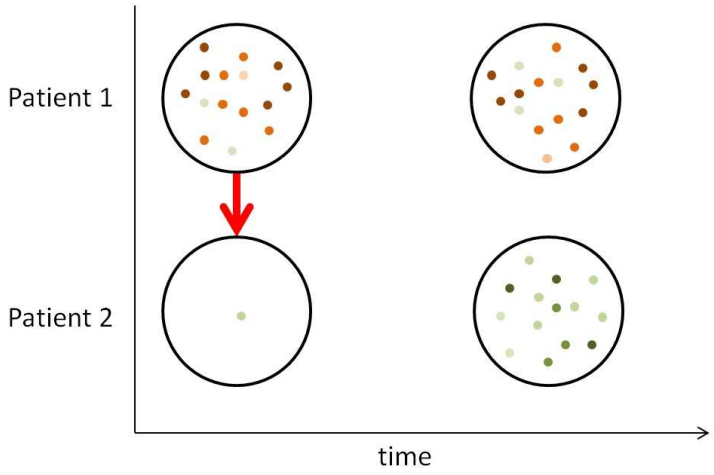

Figure 4.7: Hypothetical population bottleneck as patient 1 colonises patient 2. Coloured dots represent genetic diversity within patients (black circles). Patient 2 is colonised by one bacterial strain, which subsequently grows and diversifies into a different population to that seen in patient 1.

contribution of a genetic change as

$$\begin{aligned} P(d_{tv}, d_{ts}, d_{del} | p_{ts}, p_{tv}, p_{del}) = & (p_{ts}/L)^{d_{ts}} (1 - p_{ts}/L)^{L - d_{ts}} \\ & \cdot (p_{tv}/L)^{d_{tv}} (1 - p_{tv}/L)^{L - d_{tv}} \\ & \cdot p_{del}^{d_{del}} (1 - p_{del})^{1 - d_{del}}, \end{aligned}$$

where L is the genome length, p_{ts} and p_{tv} are the expected numbers of transitions and transversions generated in an infection event, and p_{del} is the probability of a deletion occurring during an infection event. This is a more complex model of mutation than simply counting the number of SNPs, allowing certain mutations to occur at different probabilities. This means that isolates separated by few SNPs may potentially be considered genetically distant by this measure, due to the accumulation of less probable mutations.

4.5.2.3 Genetic diversity between transmission chains

We now consider the issue of genetic diversity between 'unrelated' strains — that is, strains which are not assumed to be part of the same within-ward transmission chain. Unlike many epidemic studies, we do not assume that all cases have a common origin, since new, and possibly unrelated, cases enter the hospital over time. We now consider two possibilities to account for this diversity.

The simplest method to do this would be to assume that all distances between unrelated strains are drawn from the same distribution, representing 'global' variation. However, figures 4.2 and 4.3 suggest that this may not be appropriate — it can be seen that many patients who have a positive first swab appear to be colonised with a very similar strain. This indicates two possibilities — either these strains are part of the same transmission chain, becoming colonised on the first day or two in the ICU, or that they are positive on admission: unrelated, but genetically similar. The second possibility suggests that global genetic variation is approximately bimodal, in which unrelated strains are genetically similar (strains are of the same type), or distant (different type).

We considered introducing an additional layer of hierarchy to account for grouping of similar strains. There is evidence to suspect that randomly-selected organisms would be clustered into genetically similar types. Fraser et al., in their discussion of the classification of bacterial species, described how evolutionary and environmental forces result in the clustering of related organisms [217].

4.5.3 Model framework

Our aim was to use the collected genetic data to inform our estimate of who colonised whom, and construct the transmission network. With no genetic information or any other indicators of transmission route, it is equally likely that any of the $C(t)$ colonised patients present at time t are the source of a colonisation on day t. We proposed that the probability that a transmission occurred between patients i and j is dependent on the genetic distances observed for the individuals in question. The principle behind our method is quite simple: patients colonised by genetically similar strains are more likely to belong to the same transmission chain than those colonised by dissimilar strains.

As in previous chapters, we supposed that patients are positive on admission to the hospital with probability p. Any given susceptible patient is subject to a transmission rate of $q(t) = \beta C(t)$ at time t, that is, patients are homogeneous in terms of susceptibility and propensity to transmit. We worked in discrete time, and assumed that the colonised population on day t, $C(t)$, includes present patients colonised prior to day t, and those positive on admission. The probability of a given susceptible patient acquiring MRSA from a given colonised patient on day t is

$$(1 - \exp(-q(t))/C(t).$$

Under the model described here, a colonised patient is screened MRSA positive with

probability z. Negative patients are always screened negative. A selection of positive isolates are sequenced, generating a whole genome sequence. We suppose that when an isolate is sequenced, a set of genetic distances are generated. The ith isolate to be sequenced generates $i-1$ genetic distances. Each genetic distance is drawn from a probability distribution, which depends on the relationship between the individuals (eg. within-host, same transmission chain, etc.) from whom they have been collected. The first sequenced isolate generates no distances, since there are no previous isolates.

In this analysis, we considered the following two models, which both aim to describe the genetic diversity arising from within-host dynamics, importation and transmission.

1. **Importation structure model.** This model specifies that each sequence belongs to a group, where groups contain genetically similar sequences. In contrast to the first modelling approach described in section 4.4, the MRSA groupings and the number of groups are not pre-assigned. We supposed that any pair of isolates has a genetic distance which is drawn from one of two possible distributions;

$$P(\Psi_{i,j}=x)=\begin{cases}\mu(1-\mu)^x & \text{if } i \text{ and } j \text{ are same type;}\\ \mu_G(1-\mu_G)^x & \text{otherwise,}\end{cases}$$

   where $\mu, \mu_G \in [0,1]$, and x is an integer value taking a value between zero and the length of the genome, L. It was assumed that a patient acquires the same MRSA type as their source. With probability c, a newly imported sequence is assumed to belong to an existing group ('clustered'), otherwise, the strain is considered new, and not similar to any previously observed strain. Let g_j to be the MRSA group to which patient j's carried strain belongs. If an imported strain is considered new, and not clustered, we set $g_j = j$. Under this model, any pair of isolates taken from patients within the same transmission chain have the same expected genetic distance. The parameter vector θ is defined to be $\{p, z, \beta, \mu, \mu_G, c\}$.

2. **Transmission chain diversity model.** This model allows genetic diversity to accumulate as transmission events occur. We supposed that

$$P(\Psi_{i,j}=x)=\begin{cases}\mu\gamma^{t(r(i),r(j))}(1-\mu\gamma^{t(r(i),r(j))})^x & \text{if } i \text{ and } j \text{ are in same tree;}\\ \mu_G(1-\mu_G)^x & \text{otherwise,}\end{cases}$$

   where γ, the transmission diversity factor, takes a positive value; $\gamma < 1$ indicates an increasing diversity associated with transmission events. $r(k)$ is the patient from whom the kth isolate was collected. Under this model, strains which do not

belong to the same transmission chain are considered unrelated. As such, two imported strains are necessarily unrelated. The parameter vector θ is defined in this model to be $\{p, z, \beta, \mu, \mu_G, \gamma\}$.

For both models, we have adopted a geometric distribution to describe to genetic distances between isolates. The geometric distribution has previously been proposed to model the accumulation of SNPs [218]; alternatively, a Poisson distribution has also been suggested [152]. For two isolates taken from an individual, or individuals in the same transmission chain, it is reasonable to expect little genetic difference, provided the difference in time is not high. A decreasing probability mass function for number of SNPs seems reasonable to describe such observations. Across greater time intervals, a Poisson distribution may be more suitable, as we might expect a greater degree of change, and the probability of the isolates being identical reduces. The distribution of genetic distances between unrelated strains is difficult to determine, and depends on the sample. Uniform, or bimodal (genetically similar, or dissimilar) distributions could be used. We used a geometric distribution, which is fairly flat for a large expected value, but places a higher probability density on smaller values. As unrelated organisms might be expected to belong to genetically similar clusters [217], we believe this is appropriate.

We considered the likelihood of observing the genetic distance matrix, Ψ, and screening results, X. We believe this is a more natural framework to estimate the transmission network than considering the likelihood of observing the set of sequences themselves, Y, due to the complexity of working with the very small probabilities of observing any particular sequence. This issue is discussed in further detail in section 4.7.3.

The likelihood function may be expressed as

$$\pi(X, \Psi|\theta) = \sum_T \pi(X, \Psi|T, \theta)\pi(T|\theta), \tag{4.5.1}$$

where $T = \{t^c, \phi, s, g\}$ is the set of unobserved data which completely specify the unobserved transmission dynamics. In addition, we condition on Z, a set of observed data which we do not incorporate directly in the stochastic model, consisting of admission, discharge and screening times, and population levels at time 0. For convenience, this is excluded from notation. The component $\pi(X, \Psi|T, \theta)$ is the probability of observing the screening and genetic data, given colonisation times and sources. This accounts for the sensitivity of the swab test, and the probabilities of observing particular genetic distances, given the network structure.

The joint conditional likelihood for the importation structure model $\pi(X, \Psi|T, \theta)$, the first component in equation (4.5.1), can be written as

$$\begin{aligned}\pi(X, \Psi|T, \theta) = z^{TP(X)}(1-z)^{FN(X,T)} \\ \cdot \prod_{j=2}^{n_s}\prod_{i=1}^{j}\Big[\underbrace{\mathbf{1}_{g_i=g_j}\mu(1-\mu)^{\Psi_{i,j}}}_{\text{Same type}} \\ + \underbrace{\mathbf{1}_{g_i\neq g_j}\mu_G(1-\mu_G)^{\Psi_{i,j}}}_{\text{Different type}}\Big]\end{aligned}$$

where $TP(X)$ and $FN(X,T)$ are the number of true positive and false negative screening results, given swab results X and inferred augmented data T. The MRSA group to which an individual j belongs is denoted g_j. Similarly, the likelihood component for the transmission chain diversity model is

$$\begin{aligned}\pi(X, \Psi|T, \theta) &= P(\text{observed swab results and sequences } |\text{inferred tree, } \theta) \\ &= z^{TP(X)}(1-z)^{FN(X,T)} \\ &\cdot \prod_{j=2}^{n_s}\prod_{i=1}^{j}\Big[\underbrace{\mathbf{1}_{t(r(i),r(j))=0}\mu(1-\mu)^{\Psi_{i,j}}}_{\text{Within-patient}} \\ &+ \underbrace{\mathbf{1}_{0<t(r(i),r(j))<\infty}\mu\gamma^{t(r(i),r(j))}(1-\mu\gamma^{t(r(i),r(j))})^{\Psi_{i,j}}}_{\text{Same transmission chain}} \\ &+ \underbrace{\mathbf{1}_{t(r(i),r(j))=\infty}\mu_G(1-\mu_G)^{\Psi_{i,j}}}_{\text{Unrelated sequences}}\Big]\end{aligned}$$

The second component of equation (4.5.1), $\pi(T|\theta)$, is the probability of a particular set of colonisation times and sources, given the model parameters θ. For the importation structure model, this is defined as

$$\begin{aligned}\pi(T|\theta) &= P(\text{ inferred transmission dynamics } |\,\theta) \\ &= p^{\sum_i \phi_i}(1-p)^{n-\sum_i\phi_i}c^{n_c}(1-c)^{\sum_i\phi_i-n_c}\prod_{i=1}^{n}\Big[\mathbf{1}_{t_i^c=t_i^a}+\mathbf{1}_{t_i^c\neq t_i^a}\exp\big(-\sum_{t=t_i^a}^{\min(t_i^c-1,t_i^d)}\beta C(t)\big)\Big] \\ &\cdot \prod_{\substack{j:t_j^c\neq\infty \\ \phi_j=0}}(1-e^{-\beta C(t_j^c)}),\end{aligned}$$

where n_c is the number of importations belonging to a cluster. The component for the transmission diversity model excludes the terms involving c, but is otherwise identical.

The importation structure model requires the estimation of c, the clustering parameter. The full conditional distribution for c may be derived as

$$\pi(c|\theta_{-c}, A, X) \propto c^{n_c}(1-c)^{\sum_i \phi_i - n_c}\pi(c),$$

where $\sum_i \phi_i$ is the number of importations according to current status of the augmented data A, and $\pi(c)$ is the prior density of c. If we assign c a Beta(α_c, β_c) distribution *a priori*, it follows that c may be sampled directly from the Beta($\alpha_c + n_c$, $\beta_c + \sum_i \phi_i - n_c$) distribution using a Gibbs step. In a similar fashion, p and z may be updated with a Gibbs step. All other parameters are updated using Metropolis-Hastings steps.

4.5.4 Data augmentation

Since the full transmission process is typically unobserved, a data-augmented MCMC process was used. Colonisation times were inferred, as in the algorithm described in chapter 2, but in addition, we sampled over the infection network T by inferring the source of colonisation s_j for each carrier j. The data augmentation process allows patients with no observed sequences to be colonised. This means that we need to know how genetically distant the bacteria colonising a proposed carrier (j, say) are to all observed sequenced isolates. This allows a probability to be placed on transmission to, and from, this individual. In order to do this, one 'phantom observation' is created for this individual, creating a new row (or column) of the genetic distance matrix Ψ, which we denote Ψ^c_j, when we propose to add a colonisation. This incorporates the uncertainty of unobserved colonisations to estimates of genetic diversity (μ and μ_G). Probability mass functions $m(\cdot)$ and $m_G(\cdot)$ are defined, which are used to generate distances from this imputed sequence to isolates in the same group, and different groups, respectively. Further, we define $Y_{\text{ext}}(t) = \{y_{i,1} : t^a_i < t,\, s_i = 0\}$ be the set of observed imported sequences prior to time t.

4.5.4.1 Importation structure model

We describe here the data augmentation step for the importation structure model, where the genetic distance between strains depends on their assigned type. Due to the need to classify importations by MRSA type (g), the data augmentation step is more complex than for the transmission chain diversity model. The aim of the data augmentation process is to sample over the set of missing data $T = \{s, g, t^c, \phi, \Psi^c\}$, that is, the

set of sources s, MRSA groups g, colonisation times t^c, admission statuses ϕ, and a set of unobserved genetic distances, Ψ^c.

At each iteration, a new dataset $T^* = \{s^*, g^*, t^{c*}, \phi^*, \Psi^{c*}\}$ is proposed. Any patient who has a colonisation added by the algorithm is assigned a colonisation time and source, and a set of genetic distances from all other observed and inferred isolates. Let v_s be the number of patients never screened positive, v_q be the number of patients who carry MRSA at some point during their episode (either observed, or added by the algorithm), v_a be the number of patients for whom a colonisation time has been added by the algorithm, v_0 of whom have no 'offspring'; that is, the inferred colonised patients who infect no further individuals. Finally, let v_n be the number of patients who have a positive screen, but no sequenced isolates. We define the proposal ratio $q_{A,A^*} = P(T^* \to T)/P(T \to T^*)$. At each iteration of the algorithm, one of the following moves is made with equal probability:

- **Change colonisation route/time.** Select uniformly at random one of the v_q patients (j, say) with a colonisation time. If $v_q = 0$, no move is made. With probability w, propose the patient was positive on admission ($\phi_j^* = 1$), otherwise sample a colonisation time t_j^{c*} from $\{t_j^a, \ldots, l_j\}$, where l_j is the last potential day of colonisation (the earliest from day of discharge, day of first positive screen, and first onward transmission). If an importation is proposed, then with probability w', we set g_j^* to the same group of one of the $Y_{\text{ext}}(t_j^a)$ already-observed imported patients, otherwise, set $g_j^* = j$. If an acquisition has been proposed, we then select one of the $C(t_j^{c*})$ patients already colonised on the proposed transmission day (excluding the chosen patient, if present on day t_j^{c*}) to be the source of colonisation. If there are no other colonised patients on this day, the move is rejected. We define q_{T,T^*} according to the following table, where the row denotes the current state, and the column is the proposed state:

	Acquisition	Importation ($g_j^* \neq j$)	Importation ($g_j^* = j$)
Acquisition	$\frac{C(t_j^{c*})}{C(t_j^c)}$	$\frac{\lvert Y_{\text{ext}}(t_j^a)\rvert(1-w)}{ww'(l_j-t_j^a+1)C(t_j^c)}$	$\frac{1-w}{w(1-w')(l_j-t_j^a+1)C(t_j^c)}$
Importation ($g_j \neq j$)	$\frac{ww'(l_j-t_j^a+1)C(t_j^{c*})}{\lvert Y_{\text{ext}}(t_j^a)\rvert(1-w)}$	1	$\frac{w'}{\lvert Y_{\text{ext}}(t_j^a)\rvert(1-w')}$
Importation ($g_j = j$)	$\frac{w(1-w')(l_j-t_j^a+1)C(t_j^{c*})}{1-w}$	$\frac{(1-w')\lvert Y_{\text{ext}}(t_j^a)\rvert}{w'}$	1

- **Change genetic distances.** Select one of the v_n individuals with a positive screen, but no genetic data (j, say). If $v_n = 0$, no move is made. Update their set of $n_s + v_a$

genetic distances $\Psi^{c*}_{j,1}, \ldots, \Psi^{c*}_{j,n_s+v_a}$. These distances are drawn at random according to the probability mass function m and m_G if the sequence being compared is taken from a related or unrelated chain respectively. This move has proposal ratio

$$q_{T,T^*} = \frac{\prod_{i \neq j}(\mathbf{1}_{g_i=g_j} m(\Psi^{c}_{j,i}) + \mathbf{1}_{g_i \neq g_j} m_G(\Psi^{c}_{j,i}))}{\prod_{i \neq j}(\mathbf{1}_{g_i=g_j} m(\Psi^{c*}_{j,i}) + \mathbf{1}_{g_i \neq g_j} m_G(\Psi^{c*}_{j,i}))}.$$

- **Add colonisation.** Select at random one of the $v_s - v_a$ patients (j, say) who is currently assumed to be negative. If $v_s - v_a = 0$, no move is made. With probability w, define this patient to be an importation, otherwise, an acquisition. If an importation is proposed, set $\phi^*_j = 1$, $t^{c*}_j = t^a_j$. Now, we determine whether the proposed importation is clustered (in which case a group must be chosen) or not. With probability w', propose the sequence is clustered, and select at random one of the already-observed imported sequences $Y_{\text{ext}}(t^a_j)$, setting the proposed MRSA group g^*_j to that of the chosen sequence. If $|Y_{\text{ext}}(t^a_j)| = 0$, the move is rejected. Draw a set of $n_s + v_a$ genetic distances $\Psi^{c*}_{j,1}, \ldots, \Psi^{c*}_{j,n_s+v_a}$ from probability mass functions $m(\cdot)$ and $m_G(\cdot)$, for strains in the same group and different groups respectively.

  With probability $1 - w'$, the sequence is not clustered, so the chosen individual is assigned to a new group; $g^*_j = j$. Draw a set of $n_s + v_a$ genetic distances $\Psi^{c*}_{j,1}, \ldots, \Psi^{c*}_{j,n_s+v_a}$ from the probability mass functions $m_G(\cdot)$ to all other sequences.

  If an acquisition is proposed, then draw a colonisation time t^{c*}_j from $\{t^a_j, \ldots, t^d_j\}$. Select with equal probability a transmission source s^*_j from the $C(t^{c*}_j)$ colonised patients on that day. If there are no colonised patients on this day, no move is made. Finally, select a set of $n_s + v_a$ genetic distances, according to the relationship between the chosen patient and other colonised patients.

- **Remove colonisation.** Choose at random one of the v_0 patients who have had a colonisation time added by the data augmentation process, and are not currently assumed to be the source of infection for another individual. If $v_0 = 0$, then no move is made. Set $\phi^*_j = 0$, $t^{c*}_j = \infty$, $g^*_j = 0$ and $s^*_j = 0$.

Having established the augmented data move mechanisms, the probability ratios q_{T,T^*} for adding or removing colonisation times may be given as follows:

	Importation (clustered)	Importation (unclustered)	Acquisition
Add	$\frac{(v_s-v_a)\lvert Y_{\text{ext}}(t_j^a)\rvert}{ww'(v_0+1)M_a}$	$\frac{v_s-v_a}{w(1-w')(v_0+1)M_a}$	$\frac{(v_s-v_a)(t_j^d-t_j^a+1)C(t_j^{c*})}{(1-w)(v_0+1)M_a}$
Remove	$\frac{ww'v_0M_r}{(v_s-v_a+1)\lvert Y_{\text{ext}}(t_j^c)\rvert}$	$\frac{w(1-w')v_0M_r}{v_s-v_a+1}$	$\frac{v_0(1-w)M_r}{(t_j^d-t_j^a+1)(v_s-v_a+1)(C(t_j^c)-1)}$

where

$$M_a = \prod_{i=1}^{n_s+v_a} (\mathbf{1}_{g_i=g_j^*} m(\Psi_{j,i}^{c*}) + \mathbf{1}_{g_i\neq g_j^*} m_G(\Psi_{j,i}^{c*}))$$

and

$$M_r = \prod_{j:i\neq j} (\mathbf{1}_{g_i=g_j} m_C(\Psi_{j,i}^{c*}) + \mathbf{1}_{g_i\neq g_j} m_G(\Psi_{j,i}^{c*})).$$

Having sampled a candidate colonisation time/source, the candidate augmented dataset T^* is accepted with probability

$$\min\left(1, \frac{\pi(X,\Psi|T^*,\theta)\pi(T^*|\theta)}{\pi(X,\Psi|T,\theta)\pi(T|\theta)} q_{T,T^*}\right).$$

The proposal probability mass functions m and m_G, which are used to generate unobserved sequences related to a transmission source, an external imported strain, or the reference strain respectively, should be specified pre-analysis. Similarly, one must set w and w', the probabilities of selecting an importation, and choosing an importation cluster. These choices should not affect results, but will impact the convergence and mixing rates of the algorithm.

Performing this process over a large number of iterations will allow us to calculate the posterior probability that a particular transmission route exists; this can be calculated as the proportion of iterations for which an inferred route is made.

4.5.4.2 Transmission chain diversity model

The data augmentation process is implemented similarly for the transmission chain diversity model. The same moves are proposed, but the imputation of groupings, g, is not required. For reasons of brevity, we omit the full description of the data augmentation process for the transmission chain diversity model.

4.5.5 Modelling assumptions

Having established the model framework, we now summarise the assumptions we made to fit this model.

1. It was assumed that once a patient has become colonised, they remain so for the duration of their stay.
2. The specificity of the screening tests is 100%.
3. We did not differentiate between colonisation at different body sites; that is, we assume a swab taken at any body site is representative of an individual's status.
4. Patient readmissions were assumed to be independent to previous hospital episodes.
5. The specific biological processes occurring to generate genetic diversity were not specifically modelled. The parameters μ and μ_G represent the observed diversity, which account for diversity which may arise by sampling from a population of multiple strains, single point mutations, and homologous recombination.
6. We assumed that genetic distances are generated independently. In reality, this is not the case: for isolates A, B and C, the distance $\psi(B, C)$ is bounded by the other two distances;

$$|\psi(A, B) - \psi(A, C)| \leq \psi(B, C) \leq \min(\psi(A, B) + \psi(A, C), L),$$

where L is the total genome length. Clearly, all observed distances are generated from real DNA sequences, but imputed distances may be such that no actual sequence can conform to them. Independently-drawn genetic distances may violate these bounds, but this greatly simplifies the generation of new sequences in the data augmentation process. Interest lies in the cloud of diversity associated with within-host carriage and between-host transmission, rather than the exact composition of DNA sequences, and we do not believe violations of the above relationship for imputed sequences should affect the analysis greatly.

In this analysis, two models are considered in depth (the importation structure model, and the transmission chain diversity model). The parameters p, z and β are common to both models and have the same interpretation. In both models, we estimated μ and μ_G, which have a slightly different interpretation in each model (within and between MRSA type vs. within and between transmission chain). The importation structure model further requires the estimation of c, the clustering parameter, while the transmission chain diversity model includes g, the transmission diversity factor.

We chose uninformative prior distributions for the parameters. We assigned p, z, μ, μ_G, and c flat prior distributions on the unit interval. The parameters γ and β were given exponential distributions with rate 10^{-6} *a priori*.

4.5.6 Simulated data

In order to assess the performance of our model, we simulated epidemiological and genetic data for hospital wards according to each model. We now describe in detail how data may be simulated under either of the models described.

Patient episodes are generated with probability p of carriage on admission, and a length of stay is drawn from a Poisson distribution with mean D. Tests are generated every k calendar days, and positive patients are observed to be negative with probability $(1-z)$.

Patients positive on admission are assigned a set of genetic distances to all previously observed sequences (if applicable) which are drawn from distributions according to the relationship between isolates. For the transmission chain diversity model, genetic distances are generated by randomly drawing samples from a Geom(μ_G) distribution. For the importation structure model, an importation sequence is defined to be unclustered if no previous importation sequences have been recorded. If the sequence is not the first to be observed, the strain is defined to be clustered with probability c, otherwise, it is unclustered. For genetic distances to isolates of the same type, we draw genetic distances at random according to the Geom(μ) distribution, while for sequences in a different group, genetic distances are drawn from the Geom(μ_G) distribution.

Susceptible patients become colonised at a rate of $\beta C(t)$ at time t. Colonised patients contribute to the colonised population $C(t)$ from the day after acquisition, or the day of importation, until the day of discharge. For a newly colonised patient j, a transmission source s_j is chosen uniformly at random from the $C(t_j^c)$ positive patients present at the start of the day of colonisation. A set of genetic distances are generated according to the relationship between this patient and all previously observed patients with sequenced isolates. Under the importation structure model, distances are drawn from the Geom(μ) or Geom(μ_G) distributions, depending on whether the the isolates are of the same type, or different type respectively. Under the transmission chain diversity model, distances are drawn from the Geom$(\mu\gamma^t)$ or Geom(μ_G) distributions, depending on whether the the isolates belong to the same transmission chain (t transmission events apart), or are unrelated, respectively.

At subsequent observation times resulting in positive results, genetic distances are generated accordingly. The first observation is assigned the same distances generated for the patient's importation/acquisition. Subsequent sequenced isolates differ from pre-

vious within-host sequences by x SNPs, where $x \sim \text{Geom}(\mu)$.

4.5.7 Network distance metrics

To quantify the performance of network reconstruction, a measure to compare the true, unobserved structure to the estimated network is required. We judged a reconstructed network by two measures: firstly, the number of true transmission routes discovered, and secondly, the number of false transmission routes estimated. However, the approaches described in section 4.5.3 do not provide one estimated 'best' network, but rather a weighted network, in which each edge has an associated probability, corresponding to our estimated posterior probability that transmission occurred from one individual to another. In order to take this into account, two methods were used to measure accuracy — calculation of tree resolution, and the receiver operating characteristic (ROC) curve.

Tree resolution may be measured for a particular probability level p_L by finding the proportion of true transmission routes with a posterior probability greater than p_L. This method was used by Ypma et al. [65] to assess the performance of their transmission network reconstruction method, using simulated data. The authors plotted the resolution against the probability level for scenarios in which either genetic or spatial data were excluded, in order to demonstrate the improved accuracy of the reconstruction by including all information. While this measure shows how well true links have been discovered, it does not account for how often false links are estimated.

For $x \in [0, 1]$, the ROC curve plots the proportion of true edges against the proportion of false edges with posterior probability greater than x. The curve represents the trade-off between the correct identification of edges and the acceptance of false edges. The area under the ROC curve indicates the accuracy of the network reconstruction — an accurate reconstruction would lose very few true connections when accepting even a low false positive rate [219]. The analysis of ROC curves has been used to assess the accuracy of Bayesian networks estimating genetic regulatory interactions [220].

In order to assess the benefit of WGS data on network reconstruction, we compared the inferred network to a naïve approach, in which genetic data are ignored. In order to plot resolution or ROC curves for this case, we created a transmission network exhibiting our uncertainty of the source. For each patient j who has at least one positive swab, we set their colonisation time t_j^c to be the midpoint of the patient's first positive swab and the preceding negative swab. We then assumed that each positive patient present

at time t_j^c is equally likely to have transmitted MRSA to patient j. If patient j's first swab is positive, we supposed that with equal probability, either colonisation has been imported, or has been acquired from one of the $C(t_j^a)$ positive patients. This assumes that sensitivity of the swab test is perfect. We note that there are more sophisticated methods to estimate the transmission network with no available genetic data, based on estimated exposure times to other infective individuals and realistic estimates of z. Cooper et al. described a method to calculate the relative probabilities of acquisition from particular sources in a hospital ward, based on exposure times and MRSA type [99].

4.5.8 Single admission reproduction number

It is possible to derive estimates for a variant of the reproduction number, R_0, of MRSA in this setting from the our procedure. Cooper et al. introduced the measure $R_{a,t}$, the net single admission reproduction number, defined as the average number of secondary cases generated during a single episode, where not everyone is necessarily susceptible [99]. A similar value may be estimated from our MCMC output. By recording accepted network edges at each iteration, the mean outdegree of each node can be estimated, which represents the number of secondary cases from each colonised patient.

4.6 Results

We now present results from the methods described earlier. We firstly investigate heterogeneity in transmission rates across the Thai isolates using the grouping technique defined in section 4.4. Such multiple population models have been described and used previously; as such, we did not conduct a large-scale simulation study to assess the performance of this model. However, it is of much more interest to explore in depth the performance of our network reconstruction method (described in section 4.5). As such, we describe the results of simulation studies using this approach, before applying it to the Thai data.

4.6.1 MRSA groups

We used the isolate grouping approach to estimate the transmission rate of genetically similar strains of MRSA in the Thai ICUs. It was found that isolates in the first ICU

Transmission rates for MRSA clusters, Thai data ICU 1

Transmission parameter	1 group	2 groups	3 groups	4 groups
a_1	0.011	0.016	0.017	0.015
	(0.005,0.020)	(0.006,0.023)	(0.007,0.022)	(0.005,0.023)
a_2	—	0.0062	0.008	0.004
		(0.001,0.025)	(0.002,0.031)	(0.001,0.030)
a_3	—	—	0.018	0.027
			(0.001,0.045)	(0,0.091)
a_4	—	—	—	0.055
				(0,0.317)

Table 4.2: Posterior median estimates of the transmission parameters using the clustering approach, for up to four groups of genetically similar isolates in ICU 1. 95% equitailed credible intervals are given in parentheses.

could be grouped into at most four genetically distinct clusters with more than one member in each, but could only form three groups for ICU 2. A greater number of groups resulted in poor transmission estimates with a large amount of uncertainty.

The isolates were grouped into varying numbers of groups pre-analysis, before the data-augmented MCMC algorithm was performed to derive parameter estimates. Transmission parameter estimates for each grouping are given in tables 4.2 and 4.3.

In both wards, there was a large cluster of isolates for which the transmission rate could be estimated with relatively low uncertainty (represented by a_1 in each ICU). A great deal of uncertainty surrounds the estimates for smaller groups. With only 43 and 40 sequenced isolates in wards 1 and 2 respectively, there was little information to provide estimates of differing transmission rates.

4.6.2 Network reconstruction for simulated datasets

We now present results from the transmission network reconstruction method. Before considering the Thai data, we analysed the performance of our methods using simulated data.

We considered a baseline case of $p = 0.05$, $z = 0.85$, $\beta = 0.005$, values which we believe to be typical for hospital wards (see for example Chapter 2, [59, 126, 221]). In addition, we set $\mu = 0.03$ and $\mu_G = 0.003$ as baseline values. These values correspond to an

Transmission rates for MRSA clusters, Thai data ICU 2

Transmission parameter	1 group	2 groups	3 groups
a_1	0.007	0.010	0.009
	(0.004,0.016)	(0.004,0.021)	(0.003,0.019)
a_2	—	0.007	0.011
		(0.002,0.025)	(0,0.252)
a_3	—	—	0.22
			(0.003,0.540)

Table 4.3: Posterior median estimates of the transmission parameters using the clustering approach, for up to three groups of genetically similar isolates in ICU 2. 95% equitailed credible intervals are given in parentheses.

expected 33 and 333 SNPs for within-group and between-group isolates respectively. Within-host diversity on this scale has been demonstrated [38, 64], while between-type diversity depends largely on the sample in question — preliminary examination of the Thai data indicated 300 SNPs could be a reasonable expected genetic distance between unrelated types. For the importation structure model, we set the clustering parameter $c = 0.3$ as the default value, while the transmission diversity factor was set at $\gamma = 0.9$ in the second model. For each simulated dataset, 250 independent patient admissions were generated over 150 days, with a mean length of stay of 8 days. Screening results were taken every three days, and genetic information was generated for positive results. We investigated the impact of altering our default values on the accuracy of the estimated transmission network, as well as our posterior estimates of the parameters.

The MCMC algorithm was run for 100,000 iterations for each simulated dataset. Figure 4.8 shows an example of a transmission network simulated under baseline assumptions (top left), and our estimation of the same network (top right). The probability of an edge, or transmission route, $i \to j$ existing is estimated by the proportion of iterations for which the source of patient j is i. The ROC curve for the estimated network (bottom left), and the resolution (bottom right) are also shown, comparing the estimation to the random selection of colonised and present patients.

Perhaps unsurprisingly, it was found that higher sensitivity values resulted in more accurate network reconstruction for both models. Figure 4.9 shows the accuracy of estimated networks increases as z varies from 0.7 to 1. As false negative results occur more frequently, we get less information (or even no information) about an individ-

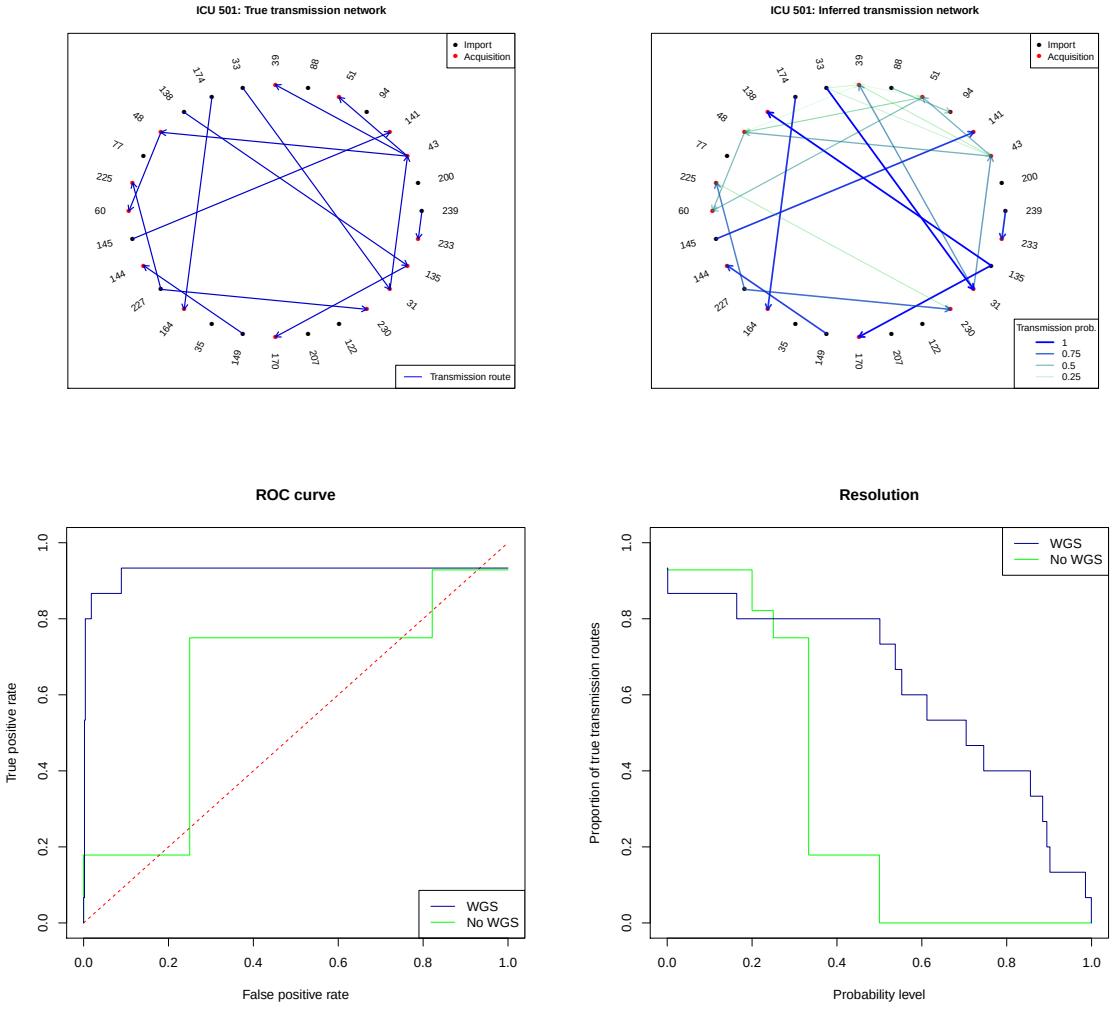

Figure 4.8: Baseline scenario true transmission network (top left) and estimated network (top right). Posterior probability of transmission routes is indicated by arrow weighting and colour. The colour of each node is on a scale depending on the probability of the individual being an importation (black) or an acquisition (red). Accuracy is shown by the ROC curve (bottom left). The dashed red line indicates theoretical random network construction. We compare the ROC curve of the estimate network (blue) to naïvely choosing a transmission source at random (green). Network resolution (bottom right) is shown, and is also compared to choosing sources randomly.

ual's MRSA type, resulting in a greater frequency of incorrect edges. Even with 100% sensitivity, some positive patients did not have a positive screen, due to colonisation and subsequent discharge occurring between swab times.

An increased transmission rate resulted in a reduction in accuracy for both models,

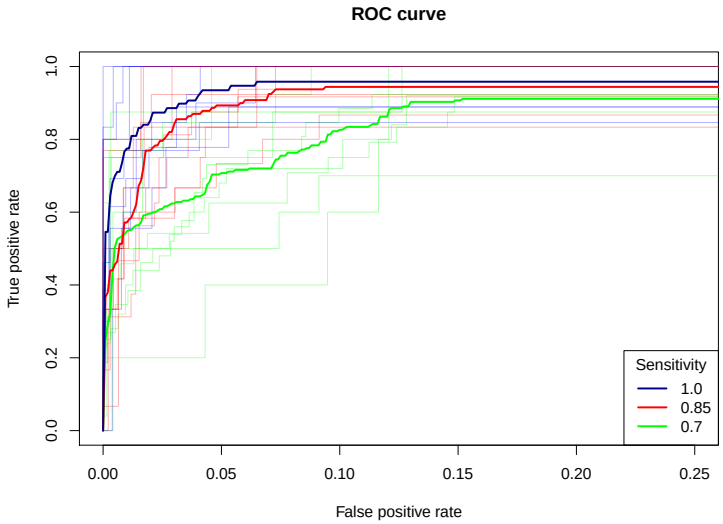

Figure 4.9: ROC curves for estimated transmission networks under various values of sensitivity for the importation structure model, restricted on the false positive rate interval $[0, 0.25]$. Ten datasets were simulated under each sensitivity level; the mean ROC curve is shown in bold.

with a slightly larger impact on the importation structure model. Clearly, when little transmission occurs, the majority of positive individuals are importations, and the network is straightforward to reconstruct. Figure 4.10 shows ROC curves for estimated networks, using data simulated under various transmission rates. Increased transmission of the same type results in the presence of several similar strains in the ward at once, which makes it more difficult to ascertain the true source of a given colonisation. Under the importation structure model, a patient assumed to acquire MRSA on a particular day will have an equal chance of having acquired the pathogen from any colonised patient within a particular group at the time. The transmission diversity model allows the transmission chain to have an impact on the expected genetic distances between two given isolates, and as such, allows greater discrimination between the set of possible sources. However, as the number of transmission events between two individuals (path length between nodes) becomes large, the expected genetic distance between isolates becomes closer to that expected under the assumption that they are unrelated ($\mu\gamma^{t(i,j)} \approx \mu_G$), potentially resulting in the incorrect conclusion that the sequences are unrelated. However, we found that as the number and length of transmission chains increases, the level of uncertainty surrounding our estimate of γ reduces.

The scale of genetic diversity within MRSA type/transmission chain, compared to that

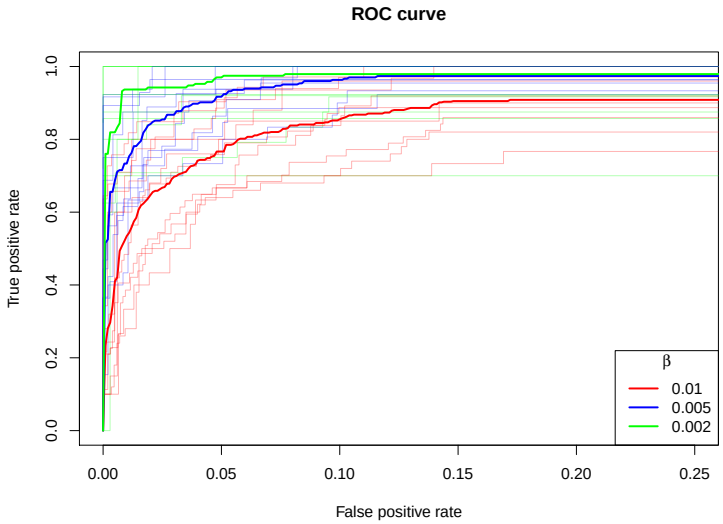

Figure 4.10: ROC curves for estimated transmission networks under various transmission rates for the importation structure model, restricted on the false positive rate interval $[0, 0.25]$. Ten datasets were simulated under each transmission rate; the mean ROC curve is shown in bold.

amongst unrelated strains, is clearly an important factor in the identification of transmission routes. By varying values of μ and μ_G, we observed the impact of this on recovering networks from simulated data. Figure 4.11 shows how accuracy reduces as μ_G approaches the same level as μ (0.03), for the transmission diversity model.

Reassuringly, we found that in almost all cases, the ROC curve for estimated transmission networks indicated a considerably better performance than the naïve approach of assigning each of the present positive patients as the source of transmission with equal probability. Only in the cases where diversity was defined to be similar for related and unrelated isolates was accuracy not improved with the incorporation of WGS data.

The importation structure model performed best when the clustering parameter c was low. When a newly observed strain is unlike anything previously observed, it is highly probable that the origin of this strain are outside of the hospital. The more frequently strains of the same type appear, the harder it becomes to differentiate between importations and acquisitions, particularly when the first observed strain is positive. In most cases, the estimate of c was associated with a large amount of uncertainty, and was poorly estimated in cases with few importations.

The performance of the transmission chain diversity model is affected by the value of γ, the transmission diversity factor. Values approaching 1 indicate that one would expect the same genetic distances from isolates taken from the same individual, as be-

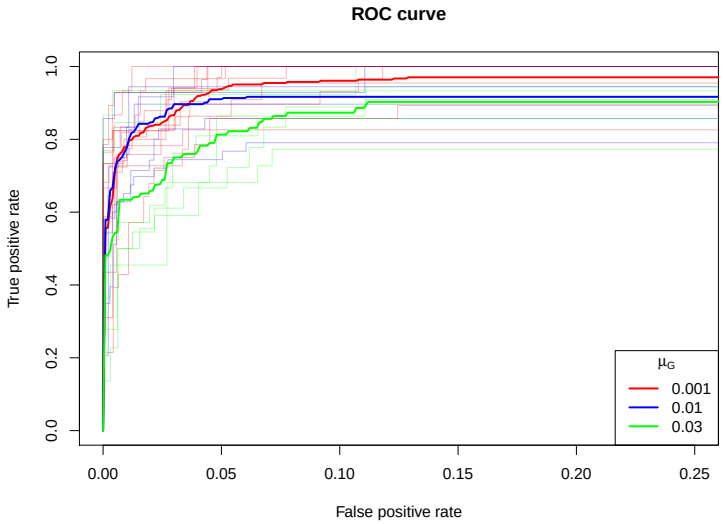

Figure 4.11: ROC curves for estimated transmission networks under various value of μ_G, genetic diversity for unrelated strains, for the transmission chain diversity model, restricted on the false positive rate interval $[0, 0.25]$. Ten datasets were simulated under each transmission rate; the mean ROC curve is shown in bold.

tween individuals in the same transmission chain. When there are multiple patients belonging to the same transmission chain present at any time, this means that genetic distances between isolates taken from these individuals are all similar, and differentiating the exact routes of transmission becomes difficult. Values of γ approaching zero indicate that there is a considerable genetic shift when a transmission event occurs, and the colony within the newly colonised individual is very different to that found in the source. This results in uncertainty as to whether these individuals are related or unrelated, as $\mu\gamma^{t(i,j)}$ and μ_G reach similar values. As such, we found that this model tended to perform poorly when γ was close to zero. Low values of γ were typically overestimated.

4.6.3 Thai data network reconstruction

We next used the real data collected from the Thai paediatric and surgical ICUs (ICU 1 and ICU 2 respectively), and attempted to estimate the transmission network in each ward. The data were analysed under both models by running the MCMC algorithm for 500,000 iterations, with 10 data augmentation steps at each iteration.

4.6.3.1 Importation structure

We first ran our analysis under the importation structure model, in which imported strains were associated with an MRSA type. Subsequent transmissions pass on MRSA of the same type. The distance between two strains was assumed to follow a geometric distribution, with parameter μ or μ_G if the strains are of the same, or of different, types respectively.

Parameter estimates are given in table 4.4. It was estimated that 6% of admissions to the surgical ICU were already colonised, compared to 16% in the paediatric ward. Estimates for test sensitivity were 74% and 83% in the two wards. Transmission rates were slightly higher in the surgical ICU; a rate corresponding to 9.5 acquisitions per 1000 patient days per colonised patient was estimated, compared to 7.7 in the paediatric ward.

Genetic diversity was estimated to be broadly similar in each ward. Any pair of MRSA isolates imputed to belong to the same group were expected to differ by 46 SNPs in ICU 1 and 38 SNPs in ICU 2. Isolates belonging to different groups were expected to differ by 402 SNPs in ICU 1, and 377 SNPs in ICU 2. There was estimated to be a greater diversity of genetic types amongst importations in ward 1 than ward 2, with the clustering parameter c estimated to be 0.2 and 0.66 respectively.

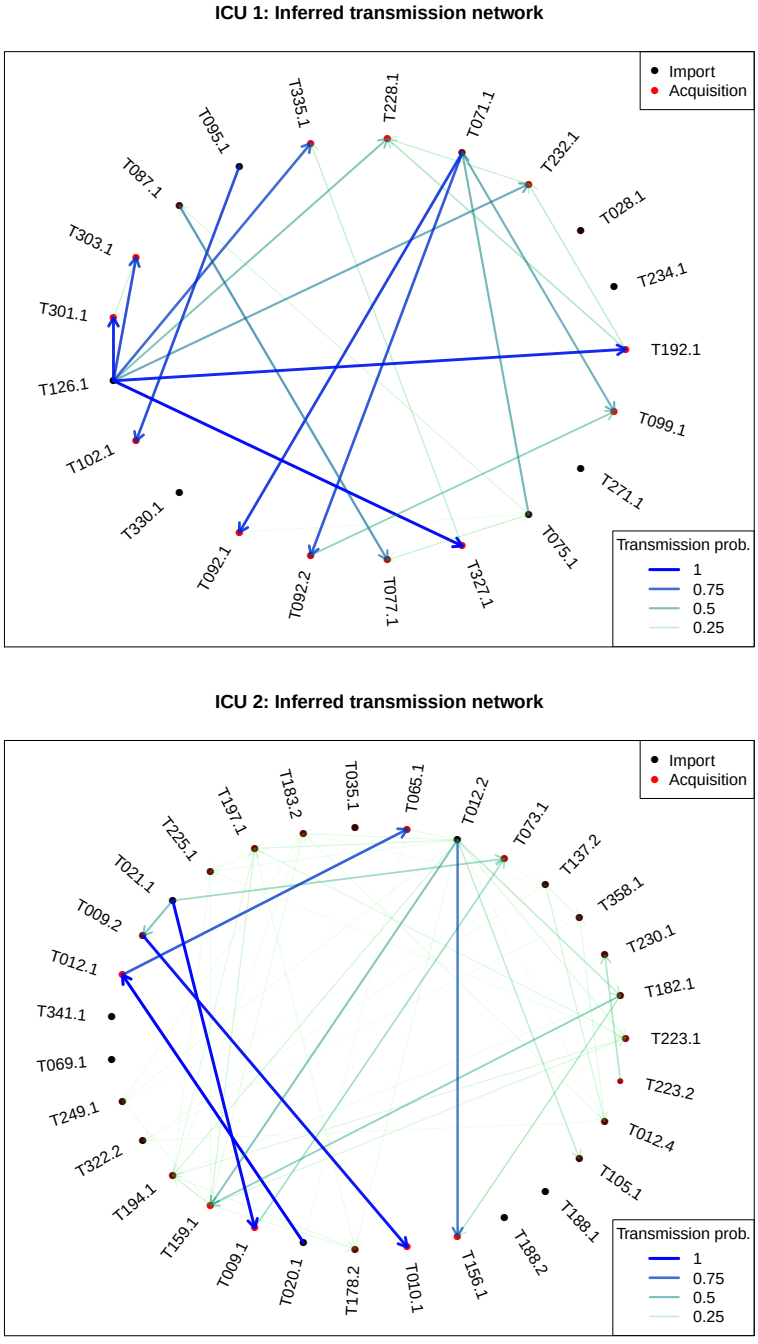

Figure 4.12: The inferred transmission networks for the surgical ICU (top) and the paediatric ICU (bottom) under the importation structure model. Colonised patients are arranged randomly on circles. The colours of the points range from red to black; the darker the colour, the more likely the patient was to have been admitted in a colonised state. Arrows represent transmission routes, arrow weighting and colour represents the posterior probability of that transmission route.

Parameter estimates under importation structure model

Parameter	ICU 1	ICU 2
Pr(importation) (p)	0.063 (0.029, 0.108)	0.164 (0.087,0.253)
Sensitivity (z)	0.735 (0.617, 0.839)	0.829 (0.746, 0.899)
Transmission rate (β)	0.0095 (0.0048, 0.0155)	0.0077 (0.0036, 0.0128)
Genetic variation (μ)	0.022 (0.019, 0.025)	0.026 (0.023,0.029)
Global genetic variation (μ_G)	0.0025 (0.0022, 0.0028)	0.0027 (0.0023, 0.0030)
Cluster parameter (c)	0.20 (0, 0.56)	0.66 (0.39, 0.86)

Table 4.4: Posterior mean estimates for model parameters under the importation structure model for both ICUs, along with 95% equitailed credible intervals.

Estimated transmission networks for ICU 1 and 2 are shown in figure 4.12. Several transmission events were linked to patient T126, in ICU 1, who was the source of an estimated five colonisations (posterior mean outdegree of node T126, 5.1 (95% CrI: 3, 6)). Five transmission events can be attributed to this individual with greater than 50% posterior probability. This patient's high transmissibility is likely to be due to their long stay in the ward (71 days), longer than any other colonised patient, although this still exceeds the expected level of transmissibility, given the estimate of β. Patient T12 was estimated to have caused more transmission events than other individuals in ward 2; this patient stayed for a total of 85 days over two episodes, and was the source of 3.7 (95% CrI: 0, 6) transmission events. However, only one transmission event can be linked to this patient with greater than 50% posterior probability.

The mean outdegree of nodes over time (secondary cases) was estimated to be 0.54 (95% CrI: 0.43, 0.62) in ICU 1, and 0.42 (0.26, 0.61) in ICU 2, suggesting a slightly higher expected number of secondary cases in ICU 1.

4.6.3.2 Transmission chain diversity

Next, an analysis of the data under the transmission chain diversity model was performed, in which strains were considered related if they belong to the same transmission chain, and unrelated if not. We supposed that the genetic distance between two isolates follows a geometric distribution with mean $1/(\mu\gamma^t)$, where t is the shortest number of edges between the two nodes associated with the sequences in the transmission network. The distance between unrelated sequences was assumed to follow a geometric distribution with mean $1/\mu_G$.

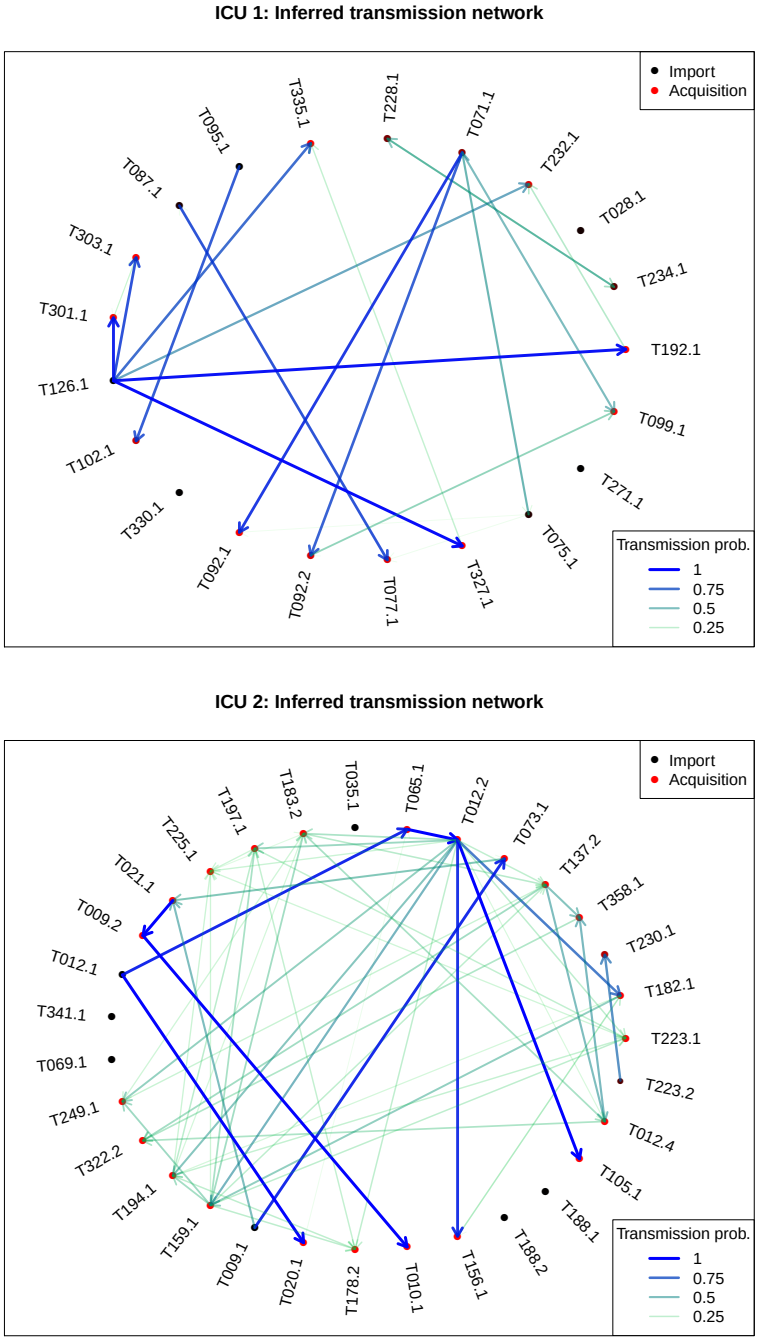

Figure 4.13: The inferred transmission networks for the surgical ICU (top) and the paediatric ICU (bottom) under the transmission chain diversity model. Colonised patients are arranged randomly on circles. The colours of the points range from red to black; the darker the colour, the more likely the patient was to have been admitted in a colonised state. Arrows represent transmission routes, arrow weighting and colour represents the posterior probability of that transmission route.

Parameter estimates for ICU 1 under this model were found to be broadly similar to those estimated for the importation structure model. We estimated that γ, the parameter describing the change in genetic diversity due to transmission events was very close to 1, indicating that between patient genetic distances differed very little from the genetic distance seen from within-patient samples. Conversely, estimates for ICU 2 varied considerably under this model. A lower rate of importation was estimated (estimated to be half of that found for the importation structure model), but the rate of acquisition was found to be higher. Possible explanations for this are discussed later. We again found that the estimate of γ was close to 1, indicating little change in diversity due to transmission events.

Parameter estimates under transmission chain diversity model

Parameter	ICU 1	ICU 2
Pr(importation) (p)	0.064 (0.032, 0.107)	0.083 (0.038,0.139)
Sensitivity (z)	0.754 (0.637, 0.856)	0.845 (0.765, 0.911)
Transmission rate (β)	0.0102 (0.0054, 0.0164)	0.0101 (0.0064, 0.0147)
Genetic variation (μ)	0.024 (0.021, 0.028)	0.029 (0.022,0.036)
Global genetic variation (μ_G)	0.0026 (0.0024, 0.0028)	0.0027 (0.0024, 0.0030)
Transmission diversity (γ)	1.03 (0.87, 1.21)	0.98 (0.92, 1.05)

Table 4.5: Posterior mean estimates for model parameters under the transmission chain diversity model for both ICUs, along with 95% equitailed credible intervals.

The estimated transmission networks for ICU 1 and 2 are shown in figure 4.13. Notably, while the estimated network for ICU 1 remained very similar under this model compared to the importation structure model (figure 4.12), the network for ICU 2 is quite distinct. This corresponds to the differences in parameter estimates under each model, and is discussed later.

We found that patients typically had a higher outdegree in ICU 2 under this model than the importation structure model, corresponding to the higher estimate for the transmission rate. For instance, patient T12 was estimated to be the source of 8.6 (95% CrI: 4, 13) transmission events, compared to 3.7 under the previous model.

Under the transmission chain diversity model, the mean outdegree of nodes was estimated to be 0.56 in ICU 1, and 0.72 in ICU 2, indicating a higher expected number of secondary cases in ICU 2.

4.6.3.3 No WGS information

Finally, we ran an analysis without using the WGS data in order to compare results and determine the impact of its incorporation. This was performed by augmenting only colonisation times rather than sources and strain types, using the same method described in chapter 2. This allowed estimates for p, z and β to be derived. Results are given in table 4.6.

Estimated parameter values for separate ICUs (no WGS data))

Parameter	ICU 1 estimates (95% CrI)	ICU 2 estimates (95% CrI)
Pr(importation) (p)	0.046 (0.008, 0.105)	0.193 (0.112,0.286)
Sensitivity (z)	0.759 (0.626, 0.874)	0.862 (0.776,0.929)
Transmission rate (β)	0.0116 (0.0059, 0.0185)	0.0071 (0.0032,0.0123)

Table 4.6: Posterior mean parameter estimates for each ICU, without the use of genetic data, along with 95% credible intervals.

Estimates of sensitivity were found to be similar to those obtained using WGS data. However, estimates of p here vary compared to those found under the transmission chain diversity model, and the importation structure model (lower for ICU 1, and higher for ICU 2). These differences are discussed in the next section.

4.7 Discussion

We have described methods to analyse the transmissibility of MRSA types, by grouping isolates according to genetic similarity. This analysis may be applied to WGS data, but also lower resolution genetic data, sufficient to partition isolates according to similarity. This method was applied to data collected from ICUs in Thailand, in order to detect differences in transmissibility.

In addition, we have provided methods to unite the analysis of genetic and epidemiological data for the transmission of MRSA in hospitals. More generally, the approaches we have used can be applied to the analysis of disease transmission where multiple importations can occur. We applied our methods to data collected from Thai ICUs, and estimated a transmission network for each ward.

4.7.1 MRSA grouping method

The MRSA grouping approach found fairly little difference in the transmissibility of genetically distinct types in the Thai data. In both ICUs, there appears to be one dominant group of isolates (see figure 4.14), which were identified as group 1 when partitioning the isolates. The estimate of the transmission rate from these groups is similar to the overall estimate of transmission, when treating all strains equally (tables 4.2 and 4.3). Other groups appear to be less clearly defined, and the uncertainty surrounding the estimates becomes very large when analysing three or more clusters. It seems likely that the limited size of the dataset means that any actual difference in transmissibility would not be estimated well. It is possible that some of the isolates outside the dominant group belonged to a less common but more transmissible type, but this was not evident due to lack of inferred transmission events.

Another concern with this approach is the assumption that genetic distance is related to differences in transmissibility, and that isolates in the same groups are equally likely to be transmitted. It is possible that a few SNPs could result in a more transmissible type, but our measure of genetic similarity may put these isolates in the same group.

4.7.2 Network reconstruction

The methods described here are a novel approach to incorporating genetic data to the study of pathogen transmission dynamics. This framework allows the simultaneous estimation of model parameters and a transmission network. These methods offer flexibility not available in previous approaches, allowing for multiple independent transmission trees, unobserved colonisation times, and imperfect observations.

Simulation studies revealed that both models perform well in most scenarios, in terms of recovering parameter values and reconstructing the transmission network. Both may be appropriate in different settings. The importation structure model may be used when it is believed that imported isolates belong to multiple genetically similar groups. The transmission chain diversity model may be used to investigate the hypothesis that a genetic shift occurs upon transmission from one individual to another.

We found that estimated parameter values differed considerably between the models when examining ICU 2. The difference stems from the designation of importation or acquisition to those individuals who are positive at their first screen. Figure 4.3 shows several patients carrying a genetically similar strain fall into this category. The impor-

tation structure model tends to assign such individuals as importations belonging to the same group, since the probability of acquiring MRSA in the first day or two was typically lower. In contrast, under the transmission chain diversity model, the probability of having such a similar strain, but being unrelated to other patients in this strain, is considered unlikely.

An analysis of the Thai data excluding genetic data was additionally performed. Clearly, it was impossible to estimate the genetic diversity or clustering parameters, but we were able to compare our estimates of p, z and β to the models utilising genetic data. While our estimates of sensitivity remained similar to those found under the transmission chain diversity model and the importation structure model, our estimate of p differed (lower for ICU 1, and higher for ICU 2). With no genetic data, the probability of being colonised on admission is largely dependent on the time of the first positive observation — if this is soon after admission, the probability of importation is relatively high. As such, many of the individuals who appear to be positive on admission in ICU 2 (see figure 4.3) are determined to be importations. However, when genetic data are available, it can be seen that many of these individuals carry a genetically similar type, lending support to the hypothesis that these patients may have acquired colonisation soon after admission.

It was assumed in our analyses that patient admission statuses (ϕ) were independent. However, a small number of patients (1 in ICU 1 and 4 in ICU 2) were readmitted to the ward, having been discharged at an earlier time. Readmission episodes were assumed to be independent, the patient being positive on admission with probability p. In reality, due to the typically lengthy carriage period for MRSA [169, 170], patients positive during their first episode are likely to still be positive upon readmission within the short study span. For this reason, we repeated our analysis, supposing that these patients could not be susceptible on admission. As expected, this resulted in an increase in the estimate of p in both wards, and a slight reduction in transmission rate β. All other parameters remained approximately the same. Other than the nodes restricted to be positive on admission, there was little change in the estimated transmission network.

We noted that there were a small number of individuals who appeared to be colonised by extremely diverse MRSA strains at different times (see figure 4.6). Since the second observed strain appears similar to other isolates present in the ward at time of observation, we believe that this within-host diversity arises from secondary colonisation,

which either coexists with, or replaces, the initial bacterial colony. These isolates are likely to increase the estimate of within-patient diversity, leading to a lower estimate for μ. This could additionally impact the estimated transmission network. The analysis was repeated, excluding any isolates which differed by 100 SNPs or more from the patient's previous isolate (three cases in ICU 1, one in ICU 2). This resulted in an increased estimate of μ (expected number of SNPs between isolates of the same type in ICU 1: 34, down from 46). There was also a slight increase in estimates for μ_G.

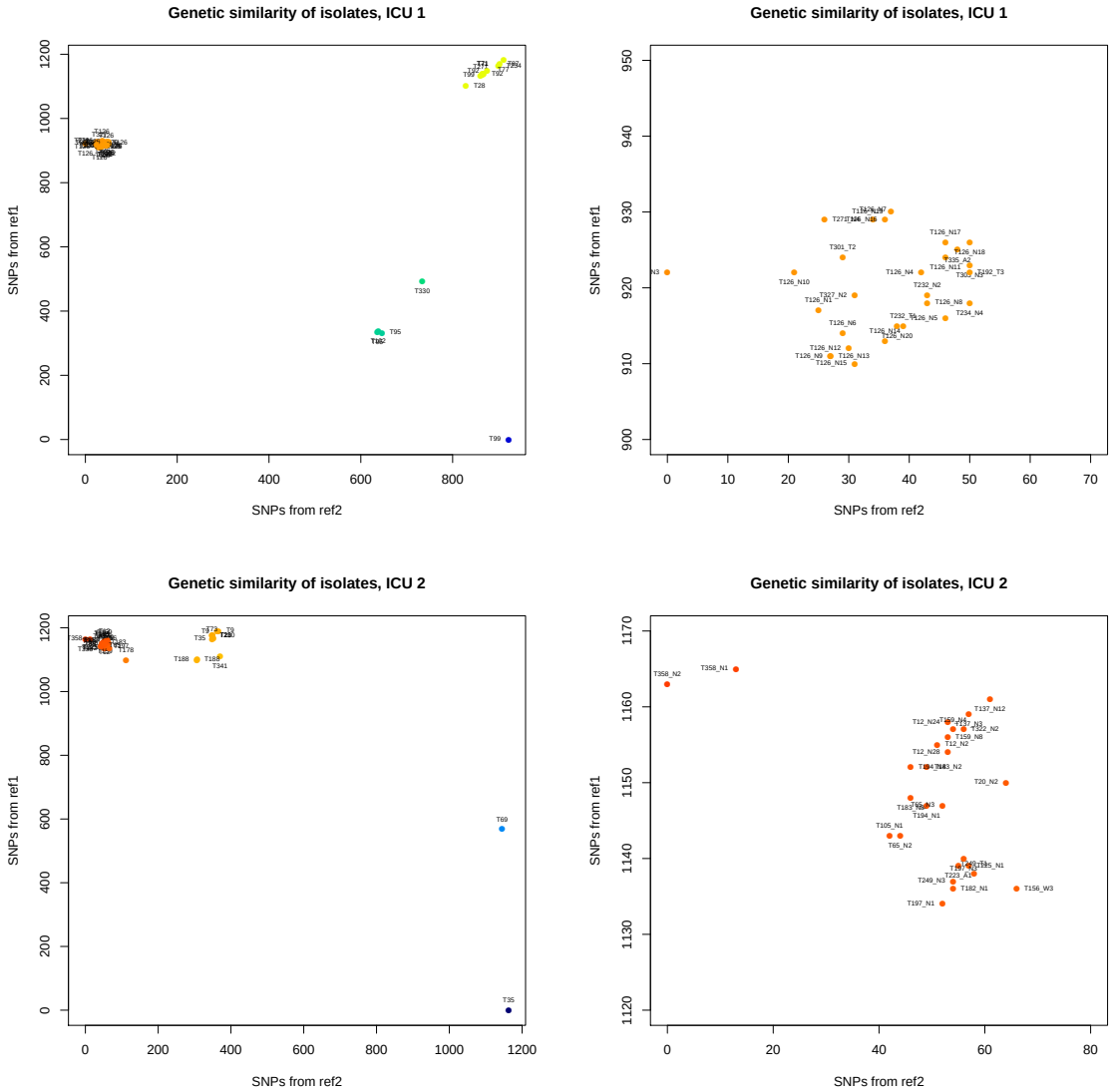

Figure 4.14: Genetic similarity of isolates sequenced in ICU 1 (top) and ICU 2 (bottom), shown in relation to two dissimilar reference strains. All strains for each ICU are shown in the left column, while the right column is zoomed in on the largest cluster in each ICU.

Our analyses indicated that there was no detectable change in genetic diversity as a result of between-patient transmission, estimating the transmission diversity factor γ to be very close to 1 in both wards (see table 4.5). However, the large amount of genetic diversity within-patient (we expected over 20 SNPs between any two within-patient isolates) might obscure any such effect.

We have assumed that related genetic distances are geometrically distributed. A histogram of observed genetic distances is given in figure 4.15. It would be of interest to explore alternative distributions to describe genetic diversity, which may be more appropriate.

Other studies have measured within-host genetic diversity over time. Young et al. studied three *S. aureus* carriers in detail, and detected 30-42 SNPs within-host [38], while Mwangi et al. observed 35 mutations from a colonised individual [64], both broadly in line with our estimates. These studies were investigating *in vivo* evolution of *S. aureus* during progression from carriage to disease, and during the development of drug resistance. Mutations were analysed in detail to determine their role in particular phases of carriage and disease. No estimates of the genetic diversity existing at any one time were made. As yet, there have been few studies on within-host genetic diversity, and these have investigated a very small sample of individuals. The Thai data had few examples of patients sequenced multiple times during carriage, and as such, it is difficult to draw many conclusions about the *in vivo* genetic behaviour or diversity of *S. aureus*. A larger scale study with multiple sequenced isolates per carrier would certainly shed more light on this.

Our aim is similar to that described by Jombart et al. in 2011 [152]. In this paper, the authors developed a network optimisation algorithm called *SeqTrack* to reconstruct disease outbreaks. The authors consider a transmission network to be a partially-observed weighted graph $\mathcal{G} = \{S, E, w\}$, where S is the set of vertices, E is the set of edges, and $w : E \to \mathbb{R}$ is a weight function which assigns a weighting to each possible ancestry. The algorithm determines the subset $B \subseteq E$ which minimises $\sum_{e \in B} w(e)$, constrained by observation times — that is, individuals may only be infected by previously-observed infectious individuals. In our setting, there is no reason to suppose that transmission events occur in the same order as positive swab results, and this approach would be inappropriate. Our methods have no such limitations, and do not attempt to recover the optimal graph, returning instead graphs of a high posterior probability, which we may use to give posterior probabilities for each edge. Furthermore, our approach allows

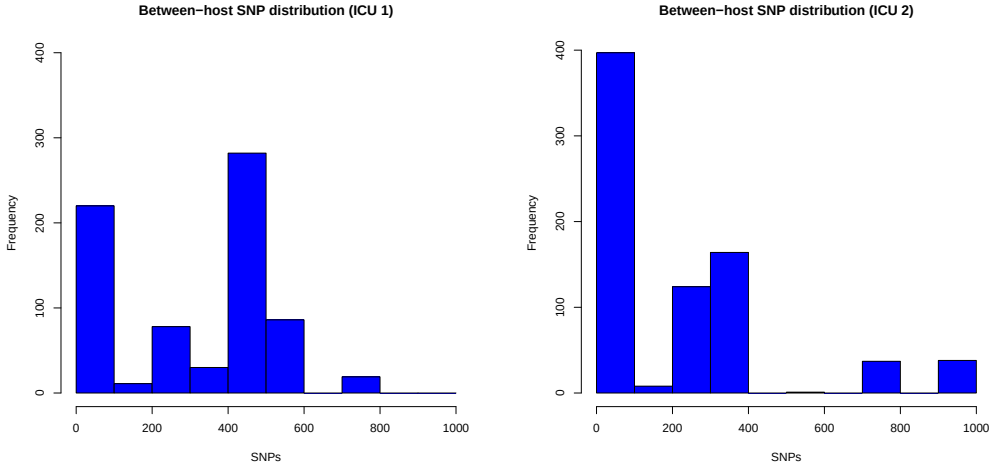

Figure 4.15: Histograms for the pairwise SNP distances recorded from ICU 1 (left) and ICU 2 (right).

individuals to have multiple strains, and within-host variation is incorporated.

Another study which aimed to reconstruct transmission networks was introduced by Ypma et al. in 2012 [65], who investigated the outbreak of avian influenza, which infected several Dutch poultry farms in 2003. The authors constructed a likelihood function with independent components representing infectiousness, genetic distance and spatial location, and use a data-augmented MCMC approach to estimate a transmission network and associated parameters, accommodating missing data. Like many other epidemic analyses, this approach assumes a single origin of disease in the population, excluding the possibility of multiple introductions. While this is certainly a realistic assumption in this case, it would not be appropriate to a hospital setting, in which introductions are frequent, and the transmission network is likely to be composed of several subtrees. Multiple introductions complicate the analysis of transmission networks, as a mechanism to differentiate between related and unrelated isolates is required. Furthermore, it is not possible to simulate data under the likelihood provided in this study, since genome sequences are utilised, but the likelihood is based only on the genetic distance between strains.

It seems intuitive that the availability of genetic data to aid the estimation of transmission routes should increase the precision of our transmission parameter estimates, particularly in a multiple population scenario, where susceptible patients experience independent colonisation pressure from different groups at different rates. However, this is not necessarily the case. We conducted a short investigation into the change in

information as transmission routes are observed. This indicated that we do not necessarily get an increase in information for particular parameters by knowing transmission routes. However, unless transmission parameters take extreme values, an overall increase in information is exhibited. Full details are given in appendix A.

4.7.3 Using nucleotide data directly

In this study, we have constructed a likelihood for observing the matrix of genetic distances, Ψ, rather than the set of genetic sequences, Y. Construction of a likelihood for the observation of Y presents some problems due to the size of the sequences. The Thai dataset contained 2591 polymorphic loci, so vectors representing sequences would have to be at least this long, but generally, could be much longer. We are interested in the change in sequences — we use the fact that sequences that are few SNPs apart are likely to be closely related, and are more likely to be present in two patients due to a transmission event, than two genetically distant sequences. Dealing with sequences, rather than genetic distances, we ask the question that, given patient i transmits to patient j, what is the probability that we observe sequence y_j from patient j, given we observed y_i from patient i? If we assume that nucleotide substitutions are equally likely, one solution is the following;

$$\begin{aligned} P(y_i|y_j, s_j = i) &= P(\psi(y_i, y_j) = x)P(x \text{ SNPs occur at loci observed}) \\ &= \mu(1-\mu)^x \frac{1}{\binom{L}{x}3^x}, \end{aligned}$$

in which L is the genome length, and we consider number of SNPs observed to be geometrically distributed with parameter μ. The 3^x term arises from assuming a nucleotide is replaced by one of the 3 other possibilities with equal probability. Clearly, this function rapidly tails off as x increases. We found that performing a data augmentation algorithm, in which we allow the source of colonisation to be updated at each step, resulted in the minimal distance network to be selected, and further proposals were almost always rejected. Since the difference in likelihood resulting from just a small number of SNPs is so great, once the closest possible source has been chosen, it is very unlikely to change. We do not believe that the minimal distance network is necessarily the correct one, and certainly not with close to 100% posterior probability.

Since we are primarily interested in the distance between sequences alone, and not the composition of sequences themselves, we believe it is convenient and reasonable to

analyse the genetic distance matrix rather than whole genome sequences themselves.

4.7.4 Future work

There are many ways in which the framework provided in this chapter may be extended. We chose simple distributions to represent the genetic diversity both within and between individuals, assuming the probability of each observed sequence was time-homogeneous. Time dependency could be introduced, allowing the expected number of SNPs to increase with time. This is approximated in the transmission chain diversity model, where it is assumed that diversity increases with transmission events (which could be assumed to be roughly proportional to time, given a relatively constant transmission rate). Furthermore, rather than simply using the number of SNPs, a more complex genetic distance function could be used, such as that used by Ypma et al. in which transitions and transversions were considered separately [65].

An extension of much interest would be to separate the effect of sampling and genetic drift. As described earlier, it seems likely that each sampled isolate is taken from a colony of genetically similar, but distinct, bacteria. By only considering one sampled isolate at each observation time, it is impossible to determine whether the distribution of strain types is changing, or whether diversity is increasing. A dataset comprising several sequenced isolates at each swab time would be ideal to study this.

We have assumed homogeneity in terms of susceptibility and transmissibility across all patients. This model may be extended to consider heterogeneous transmission rates of multiple subpopulations within the set of colonised patients. This could potentially reduce the uncertainty in estimates of intervention effectiveness, compared to the setting where no WGS data is available. This may also improve our reconstruction of the transmission network. In appendix A, the impact of observing transmission routes on the information for transmission parameters is investigated.

Incorporating mechanisms for reinfection or recombination could be of interest, and potentially important in order to reduce the impact of patients with highly genetically-distant isolates, such as those indicated in figure 4.6. In order to see the effect that these individuals have on the estimates, each could be regarded as two individuals, where each has one type of isolates. In doing this, the average genetic diversity observed within patients will not be skewed by such large distances, which in turn should reduce the uncertainty in the reconstructed networks. However, establishing a threshold above which an isolate is considered the result of a separate colonisation is likely to be

arbitrary.

A future study over a longer time period, during which isolates are collected from carriers and sequenced at regular intervals, would be of much interest. This could shed more light on the *in vivo* behaviour of the pathogen. Multiple sequenced isolates taken at each time point could clarify the extent of genetic diversity within-host at any given time, and could help to describe its dynamics over time. Isolates taken on the day of admission could reduce uncertainty around whether an individual has imported the pathogen or acquired it on the ward. As costs and resource implications fall, such a study could be feasible in the near future.

Notation used in Chapter 4	
n	Number of admissions
	Time observations
t_j^a	Admission time of patient j
t_j^d	Discharge time of patient j
$t_{j,k}^x$	Time of patient j's kth swab
$t_{j,k}^y$	Time of patient j's kth sequenced isolate
	Screening data
$x_{j,k}$	Patient j's kth swab result (positive/negative)
$y_{j,k}$	Patient j's kth sequenced isolate (DNA sequence)
ν_j	Number of swabs taken from patient j ($\nu_j \geq 0$)
ρ_j	Number of sequenced isolates for patient j ($0 \leq \rho_j \leq \nu_j$)
n_s	Total number of sequenced isolates ($n_s = \sum_i \rho_i$)
Ψ	$n_s \times n_s$ matrix of pairwise distances between all sequenced isolates
	Parameters
p	Probability of colonisation on admission
z	Test sensitivity
$a_1, \ldots, a_4$	Transmission parameters[1]
β	Transmission parameter[2,3]
μ	Within-group[2]/within-chain[3] genetic diversity
μ_G	Between-group[2]/unrelated type[3] genetic diversity
c	Clustering parameter[2]
γ	Transmission chain diversity parameter[3]
	Functions
$\psi(A, B)$	The number of SNPs between two sequences, A and B
$t(j, k)$	Shortest path from node j to node k in a transmission network[3]
$r(k)$	The patient ID associated with the kth ordered sequence[2,3]
	Latent data
ϕ_j	Admission state for patient j (1 if positive on admission, 0 otherwise)
t_j^c	Time of colonisation for patient j ($t_j^c = \infty$ if always negative)
g_j	MRSA group for patient j ($g_j = 0$ if always negative)[1,2]
s_j	Source of colonisation for patient j ($s_j = 0$ if positive on admission)[2,3]
Ψ_j^c	Set of genetic distances for a positive patient j with no sequenced isolates[2,3]

Table 4.7: [1] Notation relevant to the MRSA grouping approach (section 4.4)

[2] Notation relevant to the importation structure model (section 4.5.3)

[3] Notation relevant to the transmission chain diversity model (section 4.5.3)

CHAPTER 5

Conclusions

While numerous studies have investigated MRSA transmission in healthcare facilities, transmission dynamics are still incompletely understood, and the benefits of many interventions designed to reduce transmission are disputed. In the second chapter, we demonstrated the effectiveness of patient isolation in combination with decolonisation treatment in reducing MRSA transmission in a general ward setting. We found that the majority of transmission was associated with unisolated positive patients, underlining the importance of the detection of carriage and prompt implementation of infection control policies. No similar study has explored transmission dynamics in the general ward setting, despite the potential for such wards to act as an MRSA reservoir for the hospital as a whole. Our results provide evidence which may contribute to the discussion of cost-effective infection control measures, and may aid decisions made by policy-makers. While we provided estimates for the combination of decolonisation and isolation as a package, it is certainly of interest to explore the role of each component independently in the future. This may be achieved in a similar model framework, with data collected from a purpose-designed study.

In chapter 3, Bayesian model selection methods were considered in detail for imperfectly observed transmission models in a hospital setting. Model selection is of great importance to further knowledge of how exactly transmission dynamics work, and our systematic simulation studies contribute towards a greater understanding of the performance of reversible jump MCMC and the DIC for the particular case of transmission models. We highlighted the need to have a dataset with a large number of transmission events in order to discriminate between models. RJMCMC appeared to perform better than the DIC in our studies, and in providing posterior model probability estimates, offers an easily interpretable measure of relative model performance. These estimates

could potentially be used to calculate Bayes factors, or used as weights in a Bayesian model averaging procedure. Such procedures would be of interest when there are several models which may adequately describe the data in question, with a common value to be estimated. This investigation could be extended by considering some of the other model comparison techniques discussed in 3.2, or incorporating the prior matching techniques described in 3.6.

In chapter 4, we provided a framework with which to incorporate whole genome sequence data into the analysis of MRSA transmission. Such data are becoming increasingly available as the involved costs and processing times fall. There have been few studies utilising both genetic and epidemiological data to analyse pathogen transmission, and we believe our approach offers a new and flexible framework with which to perform such analyses. Our framework allows for multiple disease origins, multiple isolates per individual and imperfect observation, factors which complicate analysis, but have not been considered in previous studies. Simulation studies demonstrated that our methods performed well in many cases, and we estimated transmission networks based on two small datasets collected from Thai ICUs.

The models we described are relatively simple, and do not take into account many of the biological processes occurring at the molecular level which cause the observed genetic diversity. However, this provides a basis upon which more complex models may be constructed. It may be worthwhile investigating various genetic distance functions, rather than the number of SNPs, as used throughout chapter 4, or alternatives to the geometric distribution to describe genetic distance probability. The datasets we use here are insufficient to provide much insight into molecular level dynamics, but this will become feasible as larger and more detailed datasets become available. In particular, it would be of great interest to investigate the changing genetic diversity which exists within-host over time, and understand the processes behind the large amount of genetic diversity we have observed within-host. It would also be of interest to extend this model to consider the effect of antibiotics on an MRSA carrier's propensity to transmit. This could provide a greater insight into the effectiveness of such treatments. Furthermore, estimates derived under models including and excluding WGS data may be compared, to investigate the potential gain in information we discussed in appendix A.

While MRSA incidence rates are falling in many countries worldwide, it remains a major problem in resource-limited nations. Furthermore, other nosocomial pathogens are

of growing concern. Based on current trends, it is predicted that bloodstream infections caused by multiply resistant *E. coli* are likely to outnumber those caused by MRSA in the near future [222]. The threat of highly resistant gram-negative bacterial infections has become apparent in the last few years [223, 224]. The models we have described may be adapted to analyse the nosocomial transmission of such pathogens, which may be of importance as the need to evaluate infection control policies increases.

Overall, this thesis provides new insights into MRSA transmission dynamics, and a systematic study into statistical methods to compare transmission models. Finally, we have provided a framework to incorporate genetic data into transmission models. As sequence data become more abundant, the demand for such methods is likely to become ever greater, and our work provides one possible novel approach to their analysis.

References

[1] S.W. Aboelela, P.W. Stone, and E.L. Larson. Effectiveness of bundled behavioural interventions to control healthcare-associated infections: a systematic review of the literature. *Journal of Hospital Infection*, 66(2):101–108, 2007.

[2] H. Wisplinghoff, T. Bischoff, S. M. Tallent, H. Seifert, R. P. Wenzel, and M. B. Edmond. Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases form a prospective nationwide surveillance study. *Clinical Infectious Diseases*, 39(3):309–317, 2004.

[3] World Health Organisation. The burden of health care-associated infection worldwide: a summary, 2010. `http://www.who.int/gpsc/country_work/summary_20100430_en.pdf`, accessed 16th September 2012.

[4] D. Pittet, B. Allegranzi, J. Storr, S. Bagheri Nejad, G. Dziekan, A. Leotsakos, and L. Donaldson. Infection control as a major World Health Organization priority for developing countries. *Journal of Hospital Infection*, 68(4):285–292, 2008.

[5] A. K. M. Zaidi, W. C. Huskins, D. Thaver, Z. A. Bhutta, Z. Abbas, and D. A. Goldmann. Hospital-acquired neonatal infections in developing countries. *Lancet*, 365 (9465):1175–1188, 2005.

[6] R. Laxminarayan and A. Malani, editors. *Extending the cure*. Resources for the Future, Washington, DC, 2007.

[7] J. Kluytmans, A. van Belkum, and H. Verbrugh. Nasal carriage of Staphylococcus aureus: Epidemiology, underlying mechanisms, and associated risks. *Clinical Microbiology Reviews*, 10(3):505–520, 1997.

[8] R. E. O. Williams. Healthy carriage of Staphylococcus aureus: its prevalence and importance. *Bacteriology Review*, 27(1):56–71, 1963.

[9] L. A. Mermel, J. M. Cartony, P. Covington, G. Maxey, and D. Morse. Methicillin-resistant Staphylococcus aureus (MRSA) colonization at different body sites: A prospective, quantitative analysis. *Journal of Clinical Microbiology*, 49(3):1119–1121, 2011.

[10] H. F. L. Wertheim, D. C. Melles, M. C. Vos, W. van Leeuwen, A. van Belkum, H. A. Verbrugh, and J. L. Nouwen. The role of nasal carriage in Staphylococcus aureus infections. *Lancet Infectious Diseases*, 5(12):751–762, 2005.

[11] S. Harbarth, N. Liassine, S. Dharan, P. Herrault, R. Auckenthaler, and D. Pittet. Risk factors for persistent carriage of methicillin-resistant Staphylococcus aureus. *Clinical Infectious Diseases*, 31(6):1380–1385, 2000.

[12] H. F. Chambers. The changing epidemiology of Staphylococcus aureus. *Emerging Infectious Diseases*, 7(2):178–182, 2001.

[13] M. C. Enright, D. A. Robinson, G. Randle, E. J. Feil, H. Grundmann, and B. G. Spratt. The evolutionary history of methicillin-resistant Staphylococcus aureus (MRSA). *PNAS*, 99(11):7687–7692, 2002.

[14] European Centre for Disease Prevention and Control. EARS-Net database, 2012. `http://www.ecdc.europa.eu/en/activities/surveillance/EARS-Net/database`, accessed 18th July 2012.

[15] C. A. Arias and B. E. Murray. Antibiotic-resistant bugs in the 21st century - a clinical super-challenge. *New England Journal of Medicine*, 360(5):439–443, 2009.

[16] M. C. Layton, M. Perez, P. Heald, and J. E. Patterson. An outbreak of mupirocin-resistant Staphylococcus aureus on a dermatology ward associated with an environmental reservoir. *Infection Control and Hospital Epidemiology*, 14(7):369–375, 1993.

[17] M. A. Miller, A. Dascal, J. Portnoy, and J. Mendelson. Development of mupirocin resistance among methicillin-resistant Staphylococcus aureus after widespread use of nasal mupirocin ointment. *Infection Control and Hospital Epidemiology*, 17 (12):811–813, 1996.

[18] H. M. Blumberg, D. Rimland, D. J. Carroll, P. Terry, and I. K. Wachsmuth. Rapid development of ciprofloxacin resistance in methicillin-susceptible and -resistant Staphylococcus aureus. *Journal of Infectious Diseases*, 163(6):1279–1285, 1991.

[19] K. Hiramatsu, H. Hanaki, T. Ino, K. Yabuta, T. Oguri, and F. C. Tenover. Methicillin-resistant Staphyloccus aureus with reduced vancomycin susceptibility. *Journal of Antimicrobial Chemotherapy*, 40(1):135–136, 1997.

[20] P. C. Appelbaum. The emergence of vancomycin-intermediate and vancomycin-resistant Staphylococcus aureus. *Clinical Microbiology and Infection*, 12(Suppl. 1): 16–23, 2006.

[21] J.E. Coia, G.J. Duckworth, D.I. Edwards, M. Farrington, C. Fry, H. Humphreys, C. Mallaghan, and D.R. Tucker. Guidelines for the control and prevention of meticillin-resistant Staphylococcus aureus (MRSA) in healthcare facilities. *Journal of Hospital Infection*, 63:S1–S44, 2006. doi: 10.1016/j.jhin.2006.01.001.

[22] K. Hiramatsu, Y. Katayama, H. Yuzawa, and T. Ito. Molecular genetics of methicillin-resistant Staphylococcus aureus. *International Journal of Medical Microbiology*, 292(2):67–74, 2002.

[23] S. E. Cosgrove, G. Sakoulas, E. N. Perencevich, M. J. Schwaber, A. W. Karchmer, and Y. Carmeli. Comparison of mortality associated with methicillin-resistant and methicillin-susceptible Staphylococcus aureus bacteremia: A meta-analysis. *Clinical Infectious Diseases*, 36(1):53–59, 2003.

[24] S. I. Blot, K. H. Vandewoude, E. A. Hoste, and F. A. Colardyn. Outcome and attributable mortality in critically ill patients with bacteremia involving methicillin-susceptible and methicillin-resistant Staphylococcus aureus. *Archives of Internal Medicine*, 162(19):2229–2235, 2002.

[25] S. E. Cosgrove, Y. Qi, K. S. Kaye, S. Harbarth, A. W. Karchmer, and Y. Carmeli. The impact of methicillin resistance in Staphylococcus aureus bacteremia on patient outcomes: Mortality, length of stay, and hospital charges. *Infection Control and Hospital Epidemiology*, 26(2):166–174, 2005.

[26] E. M. Graffunder and R. A. Venezia. Risk factors associated with nosocomial methicillin-resistant Staphyloccus aureus (MRSA) infection including previous use of antimicrobials. *Journal of Antimicrobial Chemotherapy*, 49(6):999–1005, 2002.

[27] B. S. Cooper, G. F. Medley, and G. M. Scott. Preliminary analysis of the transmission dynamics of nosocomial infections: stochastic and management effects. *Journal of Hospital Infection*, 43(2):131–147, 1999.

[28] D. Pittet, B. Allegranzi, H. Sax, S. Dharan, C. L. Pessoa-Silva, L. Donaldson, and J. M. Boyce. Evidence-based model for hand transmission during patient care and the role of improved practices. *The Lancet Infectious Diseases*, 6(10):641–652, 2006.

[29] W. C. Albrich and S. Harbarth. Health-care workers: source, vector or victim of MRSA? *Lancet Infectious Diseases*, 8(5):389–301, 2008.

[30] R. M. Klevens, M. A. Morrison, J. Nadle, S. Petit, K. Gershman, S. Ray, L. H. Harrison, R. Lynfield, G. Dumyati, J. M. Townes, A. S. Craig, E. R. Zell, G. E. Fosheim, L. K. McDougal, R. B. Carey, and S. K. Fridkin. Invasive methicillin-resistant Staphylococcus aureus infections in the United States. *Journal of the American Medical Association*, 298(15):1763–1771, 2007.

[31] J. Marschall and K. Mühlemann. Duration of methicillin-resistant Staphylococcus aureus, according to risk factors for acquisition. *Infection Control and Hospital Epidemiology*, 27(11):1206–1212, 2006.

[32] S. J. van Hal, S. O. Jensen, V. L. Vaska, B. A. Espedido D. L. Paterson, and I. B. Gosbell. Predictors of mortality in Staphylococcus aureus bacteremia. *Clinical Microbiology Reviews*, 25(2):362–386, 2012.

[33] J. A. Jernigan, A. L. Pullen, L. Flowers, W. R. Jarvis, and M. Bell. Prevalence of and risk factors for colonization with methicillin-resistant Staphylococcus aureus at the time of hospital admission. *Infection Control and Hospital Epidemiology*, 24 (6):409–414, 2003.

[34] J. C. Lucet, S. Chevret, I. Durand-Zaleski, C. Chastang, and B. Régnier. Prevalence and risk factors for carriage of methicillin-resistant Staphylococcus aureus at admission to the intensive care unit: results of a multicenter study. *Archives of Internal Medicine*, 163(2):181–188, 2003.

[35] I. F. Chaberny, S. Ziesing, F. Mattner, S. Bärwolff, C. Brandt, T. Eckmanns, H. Rüden, D. Sohr, K. Weist, and P. Gastmeier. The burden of MRSA in four German university hospitals. *International Journal of Hygiene and Environmental Health*, 208(6):447–453, 2005.

[36] F. D. Lowy. Staphylococcus aureus infections. *New England Journal of Medicine*, 339(8):520–532, 1998.

[37] C. von Eiff, K. Becker, K. Machka, H. Stammer, and G. Peters. Nasal carriage as a source of Staphylococcus aureus bacteremia. *New England Journal of Medicine*, 344(1):11–16, 2001.

[38] B. C. Young, T. Golubchik, E. M. Batty, R. Fung, H. Larner-Svensson, A. A. Votintseva, R. R. Millera, H. Godwin, K. Knox, R. G. Everitt, Z. Iqbal, A. J. Rimmer, M. Cule, C. L. C. Ip, X. Didelot, R. M. Harding, P. Donnelly, T. E. Peto, D. W. Crook, R. Bowden, and D. J. Wilson. Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. *PNAS*, 109:4550–4555, 2012. doi: 10.1073/pnas.1113219109.

[39] M. E. Stryjewski and H. F. Chambers. Skin and soft-tissue infections caused by community-acquired methicillin-resistant Staphylococcus aureus. *Clinical Infectious Diseases*, 46(Suppl. 5):S368–377, 2008. doi: 10.1086/533593.

[40] B. Saïd-Salim, B. Mathema, and B. N. Kreiswirth. Community-acquired methicillin-resistant Staphylococcus aureus: An emerging pathogen. *Infection Control and Hospital Epidemiology*, 24(6):451–455, 2003.

[41] B. C. Herold, L. C. Immergluck, M. C. Maranan, D. S. Lauderdale, R. E. Gaskin, S. Boyle-Vavra, C. D. Leitch, and R. S. Daum. Community-acquired methicillin-resistant Staphylococcus aureus in children with no identified predisposing risk. *Journal of the American Medical Association*, 279(8):593–598, 1998.

[42] E. S. Yang, J. Tan, S. Eells, G. Rieg, G. Tagudar, and L. G. Miller. Body site colonization in patients with community-associated methicillin-resistant Staphylococcus aureus and other types of S. aureus skin infections. *Clinical Microbiology and Infection*, 16(5):425–431, 2010.

[43] K. Chua, F. Laurent, G. Coombs, M. L. Grayson, and B. P. Howden. Not community-associated methicillin-resistant Staphylococcus aureus (CA-MRSA)! A clinician's guide to community MRSA - its evolving antimicrobial resistance and implications for therapy. *Clinical Infectious Diseases*, 52(1):99–114, 2011.

[44] F. C. Tenover and R. V. Goering. Methicillin-resistant Staphylococcus aureus strain USA300: origin and epidemiology. *Journal of Antimicrobial Chemotherapy*, 64(3):441–446, 2009.

[45] M. B. Navarro, B. Huttner, and S. Harbarth. Methicillin-resistant Staphylococ-

cus aureus control in the 21st century: beyond the acute care hospital. *Current Opinion in Infectious Diseases*, 21:372–379, 2008.

[46] J. W. Lederer, D. Best, and V. Hendrix. A comprehensive hand hygiene approach to reducing MRSA health care-associated infections. *Joint Commission Journal on Quality and Patient Safety*, 35(4):180–185, 2009.

[47] V. Erasmus, T. J. Daha, H. Brug, J. H. Richardus, M. D. Behrendt, M. C. Vos, and E. F. van Beeck. Systematic review of studies on compliance with hand hygiene guidelines in hospital care. *Infection Control and Hospital Epidemiology*, 31(3):283–294, 2010.

[48] Healthcare associated infection: policy and guidance, 2010. `http://www.dh.gov.uk/en/Publichealth/Healthprotection/Healthcareassociatedinfection/Guidanceandpublications/DH_107012`, accessed 10th June 2011.

[49] S. Harbarth, P. M. Hawkey, F. Tenover, S. Stefani, A. Pantosti, and M. J. Struelens. Update on screening and clinical diagnosis of meticillin-resistant Staphylococcus aureus (MRSA). *International Journal of Antimicrobial Agents*, 37(2):110–117, 2011.

[50] D. Jeyaratnam, C. Whitty, K. Phillips, D. Liu, C. Orezzi, U. Ajoku, and G. French. Impact of rapid screening tests on acquisition of meticillin resistant Staphylococcus aureus: cluster randomised crossover trial. *British Medical Journal*, 336(7650): 927–930, 2008.

[51] E. Tacconelli, G. De Angelis, C. de Waure, M. A. Cataldo, G. La Torre, and R. Cauda. Rapid screening tests for meticillin-resistant Staphylococcus aureus at hospital admission: systematic review and meta-analysis. *The Lancet Infectious Diseases*, 9(9):546–554, 2009.

[52] T. Kypraios, P. D. O'Neill, S. S. Huang, S. L. Rifas-Shiman, and B. Cooper. Assessing the role of undetected colonisation and isolation precautions in reducing methicillin-resistant Staphyloccus aureus transmission in intensive care units. *BMC Infectious Diseases*, 10(29), 2010. doi: 10.1186/1471-2334-10-29.

[53] N. N. Damani. *Manual of infection control procedures*. Greenwich Medical Media Ltd., Cambridge, UK, 2nd edition, 2003.

[54] A. P. Fraise and C. Bradley, editors. *Ayliffe's Control of Healthcare-associated Infection*. Hodder Arnold, London, UK, 5th edition, 2009.

[55] M. M. L. van Rijen and J. A. J. W. Kluytmans. Costs and benefits of the MRSA search and destroy policy in a Dutch hospital. *European Journal of Clinical Microbiology and Infectious Diseases*, 28(10):1245–1252, 2009.

[56] B. Cooper, G. Medley, S. Stone, G. Duckworth, C. Kibbler, and R. Lai. Systematic review of isolation policies in the hospital management of methicillin-resistant Staphylococcus aureus. *Health Technology Assessment*, 7(39), 2003.

[57] J. A. Cepeda, T. Whitehouse, B. Cooper, J. Hails, K. Jones, F. Kwaku, L. Taylor, S. Hayman, B. Cookson, S. Shaw, C. Kibbler, M. Singer, G. Bellingan, and A. P. Wilson. Isolation of patients in single rooms or cohorts to reduce spread of MRSA in intensive-care units: prospective two-centre study. *Lancet*, 365(9456):295–304, 2005.

[58] M. Forrester and A. Pettitt. Use of stochastic epidemic modeling to quantify transmission rates of colonization with methicillin-resistant Staphylococcus aureus in an intensive care unit. *Infection Control and Hospital Epidemiology*, 26(7): 598–606, 2005.

[59] M. Forrester, A. Pettitt, and G. Gibson. Bayesian inference of hospital-acquired infectious diseases and control measures given imperfect surveillance data. *Biostatistics*, 8(2):383–401, 2007.

[60] T. Ueno and N. Masuda. Controlling nosocomial infection based on structure of hospital social networks. *Journal of Theoretical Biology*, 254(3):655–666, 2008.

[61] L. Milazzo, J. L. Bown, A. Eberst, and J. W. Crawford. Modelling of healthcare associated infections: A study on the dynamics of pathogen transmission by using an individual-based approach. *Computer Methods and Programs in Biomedicine*, 104(2):260–265, 2011.

[62] H. Grundmann, S. Hori, A. Tami, and D. J. Austin. Risk factors for the transmission of methicillin-resistant Staphylococcus aureus in an adult intensive care unit: fitting a model to the data. *Journal of Infectious Diseases*, 185(4):481–488, 2002.

[63] O. G. Pybus and A. Rambaut. Evolutionary analysis of the dynamics of viral infectionus disease. *Nature Reviews Genetics*, 10:540–550, 2009.

[64] M. M. Mwangi, S. W. Wu, Y. Zhou, K. Sieradzki, H. de Lencastre, P. Richardson, D. Bruce, E. Rubin, E. Myers, E. D. Siggia, and A. Tomasz. Tracking the in vivo

evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing. *PNAS*, 104(22):9451–9456, 2007.

[65] R. J. F. Ypma, A. M. A. Bataille, A. Stegeman, G. Koch, J. Wallinga, and W. M. van Ballegooijen. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society (Series B)*, 279: 444–450, 2012.

[66] S. R. Harris, E. J. Feil, M. T. G. Holden, M. A. Quail, E. K. Nickerson, N. Chantratita, S. Gardete, A. Tavares, N. Day, J. A. Lindsay, J. D. Edgeworth, H. de Lencastre, J. Parkhill, S. J. Peacock, and S. D. Bentley. Evolution of MRSA during hospital transmission and intercontinental spread. *Science*, 327(5964):469–474, 2010.

[67] V. L. Chan. Microbial genomes. In V. L. Chan, P. M. Sherman, and B. Bourke, editors, *Bacterial Genomes and Infectious Diseases*. Humana Press, Totowa, USA, 2006.

[68] J. Pevsner. *Bioinformatics and functional genomics*. Wiley-Blackwell, Hoboken, USA, 2009.

[69] T. Leitner and J. Albert. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *PNAS*, 96(19):10752–10757, 1999.

[70] M. A. Bracho, M. J. Gosalbes, D. Blasco, A. Moya, and F. González-Candelas. Molecular epidemiology of a Hepatitis C virus outbreak in a hemodialysis unit. *Journal of Clinical Microbiology*, 43(6):2750–2755, 2005.

[71] P. D. Walsh, R. Biek, and L. A. Real. Wave-like spread of Ebola Zaire. *PLoS Biology*, 3(11), 2005. doi: 10.1371/journal.pbio.0030371.

[72] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. Phylogenetic inference. In D. M. Hillis, C. Moritz, and B. K. Mable, editors, *Molecular Systematics*, pages 407–514. Sinauer Associates Inc., 2 edition, 1996.

[73] B. Foxman. *Molecular Tools and Infectious Disease Epidemiology*. Elsevier Academic Press, 2012.

[74] G. del Solar, R. Giraldo, M. J. Ruiz-Echevarría, M. Espinosa, and R. Daz-Orejas. Replication and control of circular bacterial plasmids. *Microbiology and Molecular Biology Reviews*, 62(2):434–464, 1998.

[75] H. Ochman, J. G. Lawrence, and E. A. Grolsman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 406:299–304, 2000.

[76] S. Dzidic and V. Bedeković. Horizontal gene transfer-emerging multidrug resistance in hospital bacteria. *Acta Pharmacologica Sinica*, 24(6):519–526, 2003.

[77] J. G. Lawrence. Horizontal and vertical gene transfer: the life history of pathogens. In W. Russell and H. Herwald, editors, *Concepts in bacterial virulence*, Contributions to Microbiology, pages 255–271. Karger, Basel, Switzerland, 2004.

[78] M. Nei and S. Kumar. *Molecular evolution and phylogenetics*. Oxford University Press, New York, USA, 2000.

[79] J. A. Lindsay and M. T. G. Holden. Staphylococcus aureus: superbug, super genome? *Trends in Microbiology*, 12(8):378–385, 2004.

[80] G. Prévost, B. Pottecher, M. Dahlet, M. Bientz, J.M. Mantz, and Y. Pimont. Pulsed field gel electrophoresis as a new epidemiological tool for monitoring methicillin-resistant Staphylococcus aureus in an intensive care unit. *Journal of Hospital Infection*, 17(4):255–269, 1991.

[81] D. Harmsen, H. Claus, W. Witte, J. Rothgänger, H. Claus, D. Turnwald, and U. Vogel. Typing of methicillin-resistant Staphylococcus aureus in a university hospital setting by using novel software for spa repeat determination and database management. *Journal of Clinical Microbiology*, 41(12):5442–5448, 2003.

[82] H. M. E. Frénay, A. E. Bunschoten, L. M. Schouls, W. J. van Leeuwen, C. M. J. E. Vandenbroucke-Grauls, J. Verhoef, and F. R. Mooi. Molecular typing of methicillin-resistant Staphylococcus aureus on the basis of protein A gene polymorphism. *European Journal of Clinical Microbiology and Infectious Diseases*, 15(1): 60–64, 1996.

[83] L. Koreen, S. V. Ramaswamy, E. A. Graviss, S. Naidich, J. M. Musser, and B. N. Kreiswirth. spa typing method for discriminating among Staphylococcus aureus isolates: Implications for use of a single marker to detect genetic micro- and macrovariation. *Journal of Clinical Microbiology*, 42(2):792–799, 2004.

[84] Staphylococcal Reference Unit. Staphylococcal reference unit, 2012. `http://www.hpa.org.uk/web/HPAweb&HPAwebStandard/Page/1200471672140`, accessed 6/6/2012.

[85] M. C. Enright, N. P. J. Day, C. E. Davies, S. J. Peacock, and B. G. Spratt. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus. *Journal of Clinical Microbiology*, 38 (3):1008–1015, 2000.

[86] D. J. Daley and J. Gani. *Epidemic modelling: an introduction*. Cambridge studies in mathematical biology. Cambridge University Press, Cambridge, UK, 2001.

[87] H. McCallum, N. Barlow, and J. Hone. How should pathogen transmisson be modelled? *TRENDS in Ecology & Evolution*, 16(6):295–300, June 2001.

[88] M. De Jong, A. Bouma, O. Diekmann, and H. Heesterbeek. Modelling transmission: mass action and beyond. *TRENDS in Ecology & Evolution*, 17(2):64, 2002.

[89] M. J. Keeling and P. Rohani. *Modeling Infectious diseases in humans and animals*. Princeton University Press, Princeton, USA, 2008.

[90] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 25(3), 2008. doi: 10.1371/journal.pmed.0050074.

[91] F. Ball, D. Mollison, and G. Scalia-Tomba. Epidemics with two levels of mixing. *The Annals of Applied Probability*, 7(1):46–89, 1997.

[92] W. O. Kermack and A. G. McKendrick. A contribution to the mathematial theory of epidemics. *Proceedings of the Royal Society*, 115(772):700–721, 1927.

[93] O. Diekmann, H. Metz, and H. Heesterbeek. The legacy of Kermack and McKendrick. In D. Mollison, editor, *Epidemic models: their structure and relation to data*, pages 95–118. Cambridge University Press, Cambridge, UK, 1995.

[94] P. O'Neill and G. Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society (Series A)*, 162(1):121–129, 1999.

[95] H. Andersson and T. Britton. *Stochastic Epidemic Models*, volume 151 of *Lecture Notes in Statistics*. Springer, New York, USA, 2000.

[96] J. G Ibrahim, M-H. Chen, and D. Sinha. *Bayesian Survival Analysis*. Springer Series in Statistics. Springer, New York, USA, 2005.

[97] O. Diekmann and J. A. P. Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. Wiley series in mathematical and computational biology. John Wiley & Sons, Ltd., New York, USA, 2000.

[98] E. Vynnycky and R. G. White. *An introduction to infectious disease modelling*. Oxford University Press, New York, USA, 2010.

[99] B. S. Cooper, T. Kypraios, R. Batra, D. Wyncoll, O. Tosas, and J. D. Edgeworth. Quantifying type-specific reproduction number for nosocomial pathogens: evidence for heightened transmission of an Asian sequence type 239 MRSA clone. *PLoS Computational Biology*, 8(4), 2012. doi: 10.1371/journal.pcbi.1002454.

[100] C. T. Kelley. *Iterative Methods for Optimization*. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, USA, 1999.

[101] E. K. P. Chong and S. H. Żak. *An Introduction to optimization*. Wiley Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc, Hoboken, USA, 2008.

[102] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B)*, 39(1):1–38, 1977.

[103] C. F. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[104] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer Series in Statistics. Springer, New York, USA, 2011.

[105] M Jamshidian and R. I Jennrich. Acceleration of the EM algorithm by using quasi-Newton methods. *Journal of the Royal Statistical Society (Series B)*, 59(3):569–587, 1997.

[106] N. G. Becker. Parameteric inference for epidemic models. *Mathematical Biosciences*, 117(1-2):239–251, 1993.

[107] A. C. Davison. *Statistical Models*. Cambridge Series in statistical and probablistic mathematics. Cambridge University Press, New York, USA, 2009.

[108] J. E. Kolassa. *Series approximation methods in statistics*, volume 88 of *Lecture Notes in statistics*. Springer, New York, USA, 2006.

[109] B. J. T. Morgan. *Applied Stochastic Modelling*. Chapman & Hall / CRC, 2009.

[110] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, New York, USA, 2008.

[111] D. Benton and K. Krishnamoorthy. Performance of the parametric bootstrap method in small sample interval estimates. *Advances and Applications in Statistics*, 2(3):269–285, 2002.

[112] C. Z. Mooney and R. D. Duval. *Bootstrapping: A nonparametric approach to statistical inference*, volume 95 of *Quantative Applications in the Social Sciences*. Sage Publications Ltd., Newbury Park, USA, 1993.

[113] T. J. DiCiccio and J. P. Romano. A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society (Series B)*, 50(3):338–354, 1988.

[114] B. Efron. Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, 72(1):45–58, 1985.

[115] T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statistical Science*, 11 (3):189–212, 1996.

[116] L. A. Moyé. *Elementary Bayesian Biostatistics*. Chapman & Hall / CRC, New York, USA, 2008.

[117] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov chain Monte Carlo in practice*, pages 1–20. Chapman & Hall / CRC, Boca Raton, USA, 1998.

[118] D. Gamerman. *Markov chain Monte Carlo — Stochastic simulation for Bayesian inference*. Texts in Statistical Science. Chapman & Hall / CRC, New York, USA, 1997.

[119] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in practice*. Interdisciplinary Statistics. Chapman & Hall / CRC, 1998.

[120] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and opitmal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1): 110–120, 1997.

[121] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[122] J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 169–193. Oxford University Press, Oxford, UK, 1992.

[123] M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434): 883–904, 1996.

[124] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[125] S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron, and P. Y. Boëlle. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine*, 23(22):3469–3487, 2004.

[126] B. S. Cooper, G. F. Medley, S. J. Bradley, and G. M. Scott. An augmented data method for the analysis of nosocomial infection data. *American Journal of Epidemiology*, 168(5):548–557, 2008.

[127] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

[128] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *PNAS*, 100(26):15324–15328, 2003.

[129] S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods. *PNAS*, 104(6):1760–1765, 2007.

[130] A. Gelman, X-L. Meng, and H. Stern. Posterior predictive assessment of model finess via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

[131] X-L. Meng. Posterior predictive p-values. *The Annals of Statistics*, 22(3):1142–1160, 1994.

[132] H. Jeffreys. Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(2):203–222, 1935.

[133] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[134] N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society (Series B)*, 70(3):589–607, 2008.

[135] T. Ando. *Bayesian model selection and statistical modeling*. Statistics: Textbooks and Monographs. Chapman & Hall / CRC, Boca Raton, USA, 2010.

[136] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of complexity and fit. *Journal of the Royal Statistical Society (Series B)*, 64: 583–639, 2002.

[137] G. Celeux, F. Forbes, C. P. Robert, and D. M. Titterington. Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–674, 2006.

[138] J. D. Perry, A. Davies, L. A. Butterworth, A. L. Hopley, A. Nicholson, and F. K. Gould. Development and evaluation of a chromogenic agar medium for methicillin-resistant Staphylococcus aureus. *Journal of Clinical Microbiology*, 42 (10):4519–4523, 2004.

[139] V. Sébille, S. Chevret, and A-J. Valleron. Modeling the spread of resistant nosocomial pathogens in an intensive-care unit. *Infection Control and Hospital Epidemiology*, 18(2):84–92, 1997.

[140] E. M. C. D'Agata, M. A. Horn, and G. F. Webb. The impact of persistent gastrointestinal colonization on the transmission dynamics of vancomycin-resistant Enterococci. *Journal of Infectious Diseases*, 185(6):766–773, 2002.

[141] H. Grundmann and B. Hellriegel. Mathematical modelling: a tool for hospital infection control. *Lancet Infectious Diseases*, 6(1):39–45, 2006.

[142] D. J. Austin, M. J. M. Bonten, R. A. Weinstein, S. Slaughter, and R. M. Anderson. Vancomycin-resistant enterococci in intensive-care hospital settings: Transmission dynamics, persistence, and the impact of infection control programs. *PNAS*, 96(12):6908–6913, 1999.

[143] I. Pelupessy, M. Bonten, and O. Diekmann. How to assess the relative importance of different colonization routes of pathogens within hospital settings. *PNAS*, 99 (8):5601–5605, 2002.

[144] B. S. Cooper, S. P. Stone, C. C. Kibbler, B. D. Cookson, J. A. Roberts, G. F. Medley, G. Duckworth, R. Lai, and S. Ebrahim. Isolation measures in the hospital management of methicillin resistant Staphylococcus aureus (MRSA): systematic review of the literature. *British Medical Journal*, 329(7465), 2004. doi: 10.1136/bmj.329.7465.533.

[145] E. McBryde, A. Pettitt, B. Cooper, and D. McElwain. Characterising an outbreak of vancomycin resistant Enterococci using hidden Markov models. *Journal of the Royal Society Interface*, 4(15):745–754, 2007.

[146] E. S. McBryde, A. N. Pettitt, and D. L . S. McElwain. A stochastic mathematical model of methicillin-resistant Staphylococcus aureus transmission in an intensive care unit: predicting the impact of interventions. *Journal of Theoretical Biology*, 245:470–481, 2006.

[147] C. C. Drovandi and A. N. Pettitt. Multivariate Markov process models for the transmission of methicillin-resistant Staphylococcus aureus in a hospital ward. *Biometrics*, 64:851–859, 2008.

[148] C. C. Drovandi and A. N. Pettitt. Using approximate Bayesian computation to estimate transmission rates of nosocomial pathogens. *Statistical Communications in Infectious Diseases*, 4(1), 2012.

[149] M. Bootsma, M. Bonten, S. Nijssen, A. Fluit, and O. Diekmann. An algorithm to estimate the importance of bacterial acquisition routes in hospital settings. *American Journal of Epidemiology*, 166(7):841–851, 2007.

[150] G. J. Gibson and E. Renshaw. Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA Journal of Mathematics Applied in Medicine and Biology*, 15(1):19–40, 1998.

[151] E. M Cottam, G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley, D. J. Paton, D. P. King, and D. T. Haydon. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society (Series B)*, 275(1637):887–895, 2008.

[152] T. Jombart, R. M. Eggo, P. J. Dodd, and F. Balloux. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390, 2011.

[153] L. E. Nicolle. Infection control programmes to control antimicrobial resistance. Technical report, World Health Organisation, 2001.

[154] Werkgroep Infectiepreventie. Policy for methicillin-restistant Staphylococcus aureus in hospitals. Technical report, Dutch Working Party on Infection Prevention, 2007. Available at: http://www.wip.nl/UK/free_content/Richtlijnen/MRSA%20hospital.pdf, accessed 15th May 2011.

[155] Department of Health. Screening for meticillin-resistant Staphylococcus aureus (MRSA) colonisation - a strategy for NHS trusts: a summary of best practice. Technical report, Department of Health, 2007. Available at: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_078134, accessed 15th October 2012.

[156] W. C. Huskins, C. M. Huckabee, N. P. O'Grady, P. Murray, H. Kopetskie, L. Zimmer, M. E. Walker, R. L. Sinkowitz-Cochran, J. A. Jernigan, M. Samore, D. Wallace, and D. A. Goldmann. Intervention to reduce transmission of resistant bacteria in intensive care. *New England Journal of Medicine*, 364(15):1407–1418, 2011.

[157] R. Jain, S. M. Kralovic, M. E. Evans, M. Ambrose, L. A. Simbartl, D. S. Obrosky, M. L. Render, R. W. Freyberg, J. A. Jernigan, R. R. Muder, L. J. Miller, and G. A. Roselle. Veterans affairs initiative to prevent methicillin-resistant Staphylococcus aureus infections. *New England Journal of Medicine*, 364(15):1419–1430, 2011.

[158] B. M. Farr and G. Bellingan. Pro/con clinical debate: Isolation precautions for all intensive care unit patients with methicillin-resistant Staphylococcus aureus colonization are essential. *Critical Care*, 8(3):153–156, 2004.

[159] C. Abad, A. Fearday, and N. Safdar. Adverse effects of isolation in hospitalised patients: a systematic review. *Journal of Hospital Infection*, 76(2):97–102, 2010.

[160] J. Vinski, M. Bertin, Z. Sun, S. M. Gordon, D. Bokar, J. Merlino, and T. G. Fraser. Impact of isolation on hospital consumer assessment of healthcare providers and systems scores: is isolation isolating? *Infection Control and Hospital Epidemiology*, 33(5):513–516, 2012.

[161] K. B. Kirkland and J. M. Weinstein. Adverse effects of contact isolation. *The Lancet*, 354(9185):1177–1178, 1999.

[162] H. L. Evans, M. M. Schaffer, M. G. Hughes, R. L. Smith, T. W. Chong, D. P. Raymond, S. J. Pelletier, T. L. Pruett, and R. G. Sawyer. Contact isolation in surgical patients: A barrier to care? *Surgery*, 134(2):180–188, 2003.

[163] S. Saint, L. A. Higgins, B. K. Nallamothu, and C. Cenoweth. Do physicians examine patients in contact isolation less frequently? A brief report. *American Journal of Infection Control*, 31(6):354–356, 2003.

[164] P. Gastmeier, F. Schwab, C. Geffers, and H. Rüden. To isolate or not to isolate? analysis of data from the German nosocomial infection surveillance system regarding the placement of patients with methicillin resistant Staphylococcus aureus in private rooms in intensive care units. *Infection Control and Hospital Epidemiology*, 25(2):109–113, 2004.

[165] F. M. MacKenzie, J. Bruce, M. J. Struelens, H. Goossens, J. Mollison, and I. M. Gould. Antimicrobial drug use and infection control practices associated with the prevalence of methicillin-resistant Staphylococcus aureus in European hospitals. *Clinical Microbiology and Infection*, 13(3):269–276, 2007.

[166] J. A. Jernigan, M. G. Titus, D. H. Gröschel, S. I. Getchell-White, and B. M. Farr. Effectiveness of contact isolation during a hospital outbreak of methicillin-resistant Staphylococcus aureus. *American Journal of Epidemiology*, 143(5):496–504, 1996.

[167] T. Gurieva, M. C. J. Bootsma, and M. J. M Bonten. Successful veterans affairs initiative to prevent methicillin-resistant Staphylococcus aureus infections revisited. *Clinical Infectious Diseases*, 54(11):1618–1620, 2012.

[168] L. G. Harris, S. J. Foster, and R. G. Richards. An introduction to Staphyloccus aureus and techniques for identifying and quantifying S. aureus adhesins in relation to biomaterials: review. *European Cells and Materials*, 4:29–60, 2002.

[169] A. Scanvic, L. Denic, S. Gaillon, P. Giry, A. Andremont, and J. Lucet. Duration of colonization by methicillin-resistant Staphylococcus aureus after hospital discharge and risk factors for prolonged carriage. *Clinical Infectious Diseases*, 15(10): 1393–1398, 2001.

[170] A. Robicsek, J. Beaumont, and L. Peterson. Duration of colonisation with methicillin-resistant Staphylococcus aureus. *Clinical Infectious Diseases*, 48(7):910–913, 2009.

[171] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. `www.R-project.org`.

[172] R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7:177–188, 1986.

[173] T. Lumley. *rmeta: Meta-analysis*, 2009. `CRAN.R-project.org/package=rmeta`.

[174] R. J. Sherertz, D. R. Reagan, K. D. Hampton, K. L. Robertson, S. A. Streed, H. M. Hoen, R. Thomas, and J. M. Gwaltney, Jr. A cloud adult: The Staphylococcus aureus-virus interaction revisited. *Annals of Internal Medicine*, 124(6):539–547, March 1996.

[175] L. Temime, L. Opatowski, Y. Pannet, C. Brun-Buisson, P.Y. Boëlle, and D. Guillemot. Peripatetic health-care workers as potential superspreaders. *PNAS*, 106(43): 18420–18425, 2009.

[176] S. Chang, A. K. Sethi, B. C. Eckstein, U. Stiefel, J. L. Cadnum, and C. J. Donskey. Skin and environmental contamination with methicillin-resistant Staphylococcus aureus among carriers identified clinically versus through active surveillance. *Clinical Infectious Diseases*, 48(10):1423–1428, 2009.

[177] M. van Rijen, M. Bonten, R. Wenzel, and J. Kluytmans. Mupirocin ointment for preventing Staphylococcus aureus infections in nasal carriers. Technical report, The Cochrane Collaboration, 2009.

[178] P. D. O'Neill and P. J. Marks. Bayesian model choice and infection route modelling in an outbreak of Norovirus. *Statistics in Medicine*, 24(13):2011–2024, 2005.

[179] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417, 1999.

[180] D. V. Lindley. A statistical paradox. *Biometrika*, 44(1):187–192, 1957.

[181] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

[182] S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

[183] S. A. Sisson. Transdimensional Markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association*, 100(471):1077–1089, 2005.

[184] D. L. Hastie and P. J. Green. Model choice using reversible jump Markov chain Monte Carlo. *Statistica Neerlandica*, 66(3):309–338, 2012.

[185] F. Al-Awadhi, M. Hurn, and C. Jennison. Improving the acceptance rate of reversible jump MCMC proposals. *Statistics and Probability Letters*, 69(2):189–198, 2004.

[186] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[187] A. P. Dempster. The direct use of likelihood for significance testing (reprint of 1974 article). *Statistics and Computing*, 7(4):247–252, 1997.

[188] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall / CRC, Boca Raton, USA, 2004.

[189] B. P. Carlin and S. Chib. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society (Series B)*, 57(3):473–484, 1995.

[190] C. Han and B. P. Carlin. Markov chain Monte Carlo methods for computing Bayes factors: a comparative review. *Journal of the American Statistical Association*, 96(455):1122–1132, 2001.

[191] P. Dellaportas, J. J. Forster, and I. Ntzoufras. On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12(1):27–36, 2002.

[192] S. J. Godsill. On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics*, 10(2):230–248, 2001.

[193] C. P. Robert, J-M Cornuet, J-M Marin, and N. S. Pillai. Lack of confidence in approximate Bayesian computation model choice. *PNAS*, 108(37):15112–15117, 2011.

[194] P. J. Neal and G. O. Roberts. Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics*, 5(2):249–261, 2004.

[195] S. P. Brooks, P. Giudici, and G. O. Roberts. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society (Series B)*, 65(1):3–55, 2003.

[196] P. J. Green. Trans-dimensional Markov chain Monte Carlo. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly structured stochastic systems*, pages 179–206. Oxford University Press, London, UK, 2003.

[197] D. L. Hastie. *Towards automatic reversible jump Markov chain Monte Carlo*. PhD thesis, University of Bristol, 2005.

[198] E. Paradis, J. Claude, and K. Strimmer. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

[199] A. W. F. Edwards and L. L. Cavalli-Sforza. The reconstruction of evolution. *Annals of Human Genetics*, 27:104–105, 1963.

[200] Z. Yang. Phylogeny reconstruction: overview. In *Computational Molecular Evolution*, Oxford Series in Ecology and Evolution, pages 71–99. Oxford University Press, Oxford, UK, 2007.

[201] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[202] S. D. W. Frost and E. M. Volz. Viral phylodynamics and the search for an 'effective number of infections'. *Philosophical Transactions of the Royal Society (Series B)*, 365 (1548):1879–1890, 2010.

[203] O. G. Pybus, M. A. Charleston, S. Gupta, A. Rambaut, E. C. Holmes, and P. H. Harvey. The epidemic behavior of the Hepatitis C virus. *Science*, 292(5525):2323–2325, 2001.

[204] A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192, 2005.

[205] E. M. Volz, S. L. Kosakovsky Pond, M. J. Ward, A. J. Leigh Brown, and S. D. W. Frost. Phylodynamics of infectious disease epidemics. *Genetics*, 183(4):1421–1430, 2009.

[206] D. A. Rasmussen, O. Ratmann, and K. Koelle. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Computational Biology*, 7 (8), 2011. doi: 10.1371/journal.pcbi.1002136.

[207] C. Cespedes, B. Sad-Salim, M. Miller, S-H. Lo, B. N. Kreiswirth, R. J. Gordon, P. Vavagiakis, R. S. Klein, and F. D. Lowy. The clonality of Staphylococcus aureus nasal carriage. *Journal of Infectious Diseases*, 191(3):444–452, 2005.

[208] K. Mongkolrattanothai, B.M. Gray, P. Mankin, A. B. Stanfill, R. H. Pearl, L. J. Wallace, and R. K. Vegunta. Simultaneous carriage of multiple genotypes of Staphylococcus aureus in children. *Journal of Medical Microbiology*, 60(3):317–322, 2011.

[209] E. J. Feil and B. G. Spratt. Recombination and the population structures of bacterial pathogens. *Annual Review of Microbiology*, 55:561–590, 2001.

[210] S. Castillo-Ramírez, P. Marttinen, J. Corander, M. Holden, Z. Gulay, H. Westh, and E. Feil. Poster sessions. studying recombination in the context of the population structure: the case of a recently emerged methicillin-resistant Staphylococcus aureus lineage. *Clinical Microbiology and Infection*, 18(S3):336, 2012.

[211] N. J. Croucher, S. R. Harris, C. Fraser, M. A. Quail, J. Burton, Mark van der Linden, L. McGee, A. von Gottberg, J. H. Song, K. S. Ko, B. Pichon, S. Baker, C. M. Parry, L. M. Lambertsen, D. Shahinas, D. R. Pillai, T. J. Mitchell, G. Dougan, A. Tomasz, K. P. Klugman, J. Parkhill, W. P. Hanage, and S. D. Bentley. Rapid pneumococcal evolution in response to clinical interventions. *Science*, 331(6016): 430–434, 2011.

[212] J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, 36(4):396–405, 1993.

[213] P. Marttinen, W. P. Hanage, N. J. Croucher, T. R. Connor, S. R. Harris, S. D. Bentley, and J. Corander. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Research*, 40(1), 2011. doi: 10.1093/nar/gkr928.

[214] X. Didelot, D. Lawson, A. Darling, and D. Falush. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics*, 186(4):1435–1449, 2010.

[215] D. J. Lawson, G. Hellenthal, S. Myers, and D. Falush. Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1), 2012. doi: 10.1371/journal.pgen.1002453.

[216] B. G. Spratt and M. C. J. Maiden. Bacterial population genetics, evolution and epidemiology. *Philosophical Transactions of the Royal Society (Series B)*, 354(1384): 701–710, 1999.

[217] C. Fraser, E. J. Alm, M. F. Polz, B. G. Spratt, and W. P. Hanage. The bacterial species challenge: making sense of genetic and ecological diversity. *Science*, 323 (5915):741–746, 2009.

[218] R. Sainudiin, A. G. Clark, and R. T Durrett. Simple models of genomic variation in human SNP density. *BMC Genomics*, 8(146), 2007.

[219] W. J. Krzanowski and D. J. Hand. *ROC Curves for Continuous Data*, volume 111 of *Monographs on Statistics and Applied Probability*. Chapman & Hall / CRC, Boca Raton, USA, 2009.

[220] D. Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19(17):2271–2282, 2003.

[221] J. V. Robotham, N. Graves, B. D. Cookson, A. G. Barnett, J. A. Wilson, J. D. Edgeworth, R. Batra, B. H. Cuthbertson, and B. S. Cooper. Screening, isolation, and decolonisation strategies in the control of meticillin resistant Staphylococcus aureus in intensive care units: cost effectiveness evaluation. *British Medical Journal*, 343:d5694, 2011. doi: 10.1136/bmj.d5694.

[222] M. E. A. de Kraker, P. G. Davey, and H. Grundmann. Mortality and hospital stay associated with resistant Staphylococcus aureus and Escherichia coli bacteremia: Estimating the burden of antibiotic resistance in Europe. *PLoS Medicine*, 8(10), 2011. doi: 10.1371/journal.pmed.1001104.

[223] P. Nordmann, L. Poirel, M. A. Toleman, and T. R. Walsh. Does broad-spectrum β-lactam resistance due to NDM-1 herald the end of the antibiotic era for treatment of infections caused by gram-negative bacteria? *Journal of Antimicrobial Chemotherapy*, 66(4):689–692, 2011.

[224] B. Li, Y. Yi, Q. Wang, P. C. Y. Woo, L. Tan, H. Jing, G. F. Gao, and C. H. Liu. Analysis of drug resistance determinants in Klebsiella pneumoniae isolates from a tertiary-care hospital in Beijing, China. *PLoS One*, 7(7), 2012. doi: 10.1371/journal.pone.0042280.

APPENDIX A

Impact of observing transmission routes on parameter estimates

While there is certainly much to learn about transmission routes by incorporating genetic data into our analysis, the effect that this additional information has on our transmission parameter estimates is less clear. While one might think that an increase in precision in parameter estimates would be likely, this is not necessarily the case. We consider here how observing transmission routes can impact transmission parameter estimates for a continuous-time, fully-observed epidemic.

Suppose we have a model with parameters $\theta = \{\theta_1, \ldots, \theta_m\}$ and a log-likelihood function $\ell(\theta)$. The observed information matrix is defined as

$$\mathcal{J}(\theta) = -\Big[\frac{\partial^2 \ell(\theta)}{\partial \theta_{i,j}^2}\Big]_{i,j \leq m},$$

the negative of the matrix of second derivatives of the log-likelihood function (Hessian matrix). The observed information of parameter θ_k evaluated at $\theta_k = x$ is then

$$\mathcal{J}(\theta_k = x) = -\Big[\frac{\partial^2 \ell(\theta; \theta_k = x)}{\partial \theta_k^2}\Big].$$

An increase in information corresponds to a reduction in uncertainty surrounding a parameter estimate.

Suppose we observe a total of n individuals in a dynamic population, with a set of entry times $t^a = \{t_1^a, \ldots, t_n^a\}$ and exit times $t^d = \{t_1^d, \ldots, t_n^d\}$. For those who become infected, we assume the infection times, t^I, are observed perfectly. Each individual who is never infected (j say) is assigned an infection time $t_j^I = \infty$. In this study, we are interested in the transmission parameters, and ignore any importation parameterisation for

simplicity. The inclusion of such a structure would not impact our conclusions about the transmission parameters. If we change entry and exit times so that all individuals are present for the duration of the study, this model is equivalent to an SIR model with deterministic recovery times.

Firstly, suppose we model the transmission rate as $q(t) = aI(t)$, where $I(t)$ is the number of infectious individuals in the population at time t. The likelihood of the infection times t^I, given the transmission model, is

$$\begin{aligned} L_1(a) &= \pi(t^I|a) \\ &= \exp\left(-\sum_{k=1}^{n}\int_{t=t_k^a}^{\min(t_k^d,t_k^I)} aI(t)\mathrm{d}t\right)\prod_{j:t_j^I\neq\infty} aI(t_{j-}^I), \end{aligned}$$

where t_{j-}^I is the time immediately prior to t_j^I. Now consider the case where all transmission routes, as well as infection times, are observed. We denote the set of transmission routes s; if individual k is infected by individual j, we have $s_k = j$. Each susceptible individual independently experiences a transmission rate of a from each infective individual.

It follows that the likelihood of our observations of transmission routes and times is then

$$\begin{aligned} L_2(a) &= \pi(t^I, s|a) \\ &= \exp\left(-\sum_{k=1}^{n}\int_{t=t_k^a}^{\min(t_k^d,t_k^I)} aI(t)\mathrm{d}t\right)\prod_{j:t_j^I\neq\infty} a \\ &= L_1(a)\prod_{j:t_j^I\neq\infty}\frac{1}{I(t_j^I)} \\ &\propto L_1(a). \end{aligned}$$

Since L_2 is proportional to L_1, we can conclude that the maximum likelihood estimate of a will be the same under both scenarios, and there will be no change in information with known transmission routes.

Now consider a situation where transmission may occur from multiple population groups $1, \ldots, G$ (for example, MRSA positive patients taking different types of antibiotics, or those under isolation precautions compared to unisolated individuals). We denote the number of infectious individuals in group k at time t as $I_k(t)$. Let $g(j)$ be the group to which infectious individual j belongs, and $g(j) = 0$ if the individual is

not infectious. We define the rate of transmission to any given susceptible individual at time t as $q(t) = \sum_{j=1}^{G} a_j I_j(t)$.

Suppose once again that we lack knowledge about the source of infection. Then the likelihood of our observations, given this transmission model, is

$$\begin{aligned} L_1(a_1, \ldots, a_g) &= \pi(t^I | a_1, \ldots, a_G) \\ &= \exp\left(- \sum_{k=1}^{n} \int_{t=t_k^a}^{\min(t_k^d, t_k^I)} q(t) \mathrm{d}t \right) \prod_{j: t_j^I \neq \infty} q(t_{j-}^I). \end{aligned}$$

We then calculate the second derivative of the log-likelihood, with respect to a given transmission parameter, a_m:

$$\begin{aligned} \ell_1(a_1, \ldots, a_G) &= \underbrace{- \sum_{k=1}^{n} \int_{t=t_k^a}^{\min(t_k^d, t_k^I)} q(t) \mathrm{d}t}_{:=K} + \sum_{j: t_j^I \neq \infty} \log(q(t_{j-}^I)) \\ \frac{\partial}{\partial a_m} \ell_1(a_1, \ldots, a_G) &= \frac{\partial K}{\partial a_m} + \sum_{j: t_j^I \neq \infty} \frac{I_m(t_{j-}^I)}{q(t_{j-}^I)} \\ \frac{\partial^2}{\partial a_m^2} \ell_1(a_1, \ldots, a_G) &= \frac{\partial^2 K}{\partial a_m^2} - \sum_{j: t_j^I \neq \infty} \frac{I_m(t_{j-}^I)^2}{q(t_{j-}^I)^2}. \end{aligned}$$

Now consider the case where all transmission times and routes are observed. Each susceptible independently experiences infective pressure from all infectious individuals, and is subject to a transmission rate of a_m from each infectious individual in group m. It follows that the likelihood of our observations of transmission routes and times is then

$$\begin{aligned} L_2(a_1, \ldots, a_G) &= \pi(t^I, s | a_1, \ldots, a_G) \\ &= \exp\left(- \sum_{k=1}^{n} \int_{t=t_k^a}^{\min(t_k^d, t_k^I)} q(t) \mathrm{d}t \right) \prod_{j: t_j^I \neq \infty} \left(\sum_{m=1}^{G} \mathbf{1}_{g(s_j)=m} a_m \right) \\ &= \exp\left(- \sum_{k=1}^{n} \int_{t=t_k^a}^{\min(t_k^d, t_k^I)} q(t) \mathrm{d}t \right) \prod_{m=1}^{G} (a_m^{N_m}), \end{aligned}$$

where N_m represents the total number of observed transmission events from individuals in group m, and $\mathbf{1}_x$ is the indicator function returning 1 when x is true, 0 otherwise. We are interested in the change in observed information by observing transmission routes, as opposed to treating each source of infection as equally likely. We derive the

second derivative of $\ell_2 = \log(L_2)$ with respect to a given transmission parameter, a_m:

$$\begin{aligned}
\ell_2(a_1, \ldots, a_G) &= K + \sum_{m=1}^{G} N_m \log(a_m) \\
\frac{\partial}{\partial a_m} \ell_2(a_1, \ldots, a_g) &= \frac{\partial K}{\partial a_m} + \frac{N_m}{a_m} \\
\frac{\partial^2}{\partial a_m^2} \ell_2(a_1, \ldots, a_g) &= \frac{\partial^2 K}{\partial a_m^2} - \frac{N_m}{a_m^2}.
\end{aligned}$$

It now follows that

$$\begin{aligned}
\frac{\partial^2}{\partial a_m^2} \ell_2(a_1, \ldots, a_G) &= \frac{\partial^2}{\partial a_m^2} \ell_1(a_1, \ldots, a_G) - \frac{N_m}{a_m^2} + \sum_{j: t_j^I \neq \infty} \frac{I_m(t_{j-}^I)^2}{q(t_{j-}^I)^2} \\
\mathcal{J}_2(a_m) &= \mathcal{J}_1(a_m) + \frac{N_m}{a_m^2} - \sum_{j: t_j^I \neq \infty} \frac{I_m(t_{j-}^I)^2}{q(t_{j-}^I)^2} \\
\mathcal{J}_2(a_m) &= \mathcal{J}_1(a_m) + \frac{1}{a_m^2} \Big(N_m - \underbrace{\sum_{j: t_j^I \neq \infty} \Big[\frac{a_m I_m(t_{j-}^I)}{q(t_{j-}^I)} \Big]^2}_{x} \Big),
\end{aligned}$$

for all $a_m > 0$, where $\mathcal{J}(a_i)$ is i-th diagonal element in the observed information matrix. Each component of the sum x takes a value in the interval $[0, 1]$, therefore $0 \leq x \leq \sum_{i=1}^{G} N_i = N$. Clearly, we do not necessarily gain information about specific transmission parameters by incorporating observed transmission routes — for instance, if $N_m = 0$, then $\mathcal{J}_2(a_m) < \mathcal{J}_1(a_m)$.

To consider the change in information for the set of parameters as a whole, we now examine the trace of the observed information matrix for both cases. In the case where we observe transmission routes, we may express the trace of the observed information matrix as

$$
\begin{aligned}
\operatorname{tr}(\mathcal{J}_2(a_1,\ldots,a_G)) &= \sum_{m=1}^{G} \mathcal{J}_2(a_m) \\
&= \sum_{m=1}^{G} \left[\mathcal{J}_1(a_m) + \frac{1}{a_m^2}\Big(N_m - \sum_{j:t_j^I \neq \infty} \Big[\frac{a_m I_m(t_{j-}^I)}{q(t_{j-}^I)} \Big]^2 \Big) \right] \\
&= \operatorname{tr}(\mathcal{J}_1(a_1,\ldots,a_G)) + \sum_{m=1}^{G} \frac{N_m}{a_m^2} - \sum_{m=1}^{g} \sum_{j:t_j^I \neq \infty} \Big[\frac{a_m I_m(t_{j-}^I)}{q(t_{j-}^I)} \Big]^2 \\
&= \operatorname{tr}(\mathcal{J}_1(a_1,\ldots,a_G)) + \sum_{m=1}^{G} \frac{N_m}{a_m^2} - \sum_{j:t_j^I \neq \infty} \underbrace{\frac{\sum_{m=1}^{g} a_m^2 I_m(t_{j-}^I)^2}{q(t_{j-}^I)^2}}_{\leq 1} \\
&\geq \operatorname{tr}(\mathcal{J}_1(a_1,\ldots,a_G)) + \sum_{m=1}^{G} \frac{N_m}{a_m^2} - \sum_{m=1}^{g} N_m,
\end{aligned}
$$

and, if we make an additional assumption that transmission parameters take values less than or equal to 1 (as is typically the case), it follows that

$$\operatorname{tr}(\mathcal{J}_2(a_1,\ldots,a_G)) \geq \operatorname{tr}(\mathcal{J}_1(a_1,\ldots,a_G)),$$

indicating an overall gain in information associated with the observation of transmission routes. This is potentially beneficial in the estimation of transmission parameters and functions of these, such as measures of intervention effectiveness, as described in section 2.4.2.