UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Modelling the Dynamics of *Streptococcus pneumoniae* Transmission in Children

Delphine Pessoa

Master of Bioinformatics and Computational Biology

2010

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Modelling the Dynamics of *Streptococcus pneumoniae* Transmission in Children

Delphine Pessoa

Master of Bioinformatics and Computational Biology

Dissertation under the supervision of

Flávio Coelho (IGC)
Octávio Paulo (FCUL)

2010

ACKNOWLEDGEMENTS

I thank to Carlota Gonçalves, Gabriela Gomes and Raquel Sá Leão, for creating this project and allowing me to be part of it. Many thanks to Gabriela Gomes, Raquel Sá Leão and Flávio Coelho for their continuous support and insights. Last but not least I thank the CMAF for the fellowship CMAF-BI-02/09 and the IGC for a fantastic place to work.

ABSTRACT

Streptococcus pneumoniae, also known as pneumococcus, is a commensal bacterium very common in the nasopharynx of young children but that can also be found in older children and adults. Carriage may lead to infection. Although this is a rare event, it has a significant impact on human health. Indeed, diseases caused by pneumococcus include infections as common as otitis media and as dangerous as pneumonia or meningitis. It is therefore important to gain more knowledge of its transmission dynamics as modulated by environmental factors and of how it is affected by host population specificities.

To characterize pneumococcus transmission dynamics in a Portuguese day-care center, data was used from a one-year longitudinal study on the state of colonization by *S. pneumoniae* in children attending a day-care center in Lisboa, Portugal [Sá-Leão et al.2008]. The data refer to 1998, before the introduction of the seven-valent pneumococcal conjugate vaccine.

A conceptual model for pneumococcus transmission was built, which considered genotype colonizations and clearances as dependent on the number of carriers, the number of non-carriers and the values of four parameters: the clearance rate μ, the within-group transmission parameter β, the community rate of acquisition κ and the between-genotypes competition parameter ϕ. Bayesian inference was used to estimate these parameters. Colonizations and clearances were modelled as Poisson processes and the joint posterior probability distributions of the model's parameters were estimated by Markov Chain Monte Carlo sampling. The number of transitions that occured in each sampling interval was counted directly from the sampled states, assuming that children did not undergo more than one transition per sampling interval.

The posterior mean for the transmission parameters were 0.5974for β, 0.0107for κ, 0.6280 for ϕ and 0.3059 for μ.

Data was simulated using the posterior estimates for these parameters from a study of Finnish DCCs [Hoti et al.2009]. Sampling this data monthly, the method was found to give biased estimations, since the asumption that children did not undergo more than one transition per sampling interval did not hold. The precision could be improved by sampling for a longer period, 30 months were used. To significantly improve the accuracy, the sampling interval needed to be extremely short, daily samples were taken from the simulation. The estimation model used was found to be impractical. Another estimation method should be used that infers the possible carriage histories consistent with the observed states.

KEYWORDS Streptococcus pneumoniae, pneumococcus, transmission, bayesian inference, MCMC, Poisson

RESUMO

Streptococcus pneumoniae, também denominado pneumococo, é uma bactéria comensal que só coloniza o homem. A aquisição desta bactéria começa pela colonização da nasofaringe, o indivíduo torna-se portador e pode transmitir a bactéria a outras pessoas através de contacto directo. Devido aos contactos frequentes e próximos entre crianças a frequentar infantários e à imaturidade do seu sistema imunitário, as crianças até aos seis anos de idade representam um grupo onde a prevalência de portadores de pneumococo é elevada [Bogaert et al.2004]. Estudos transversais realizados em crianças a frequentar infantários, mostraram que a prevalência de pneumococo nestas crianças pode ser superior a 70% [Mato et al.2005]. Embora a colonização seja geralmente assintomática, pode, por vezes, evoluir para infecção. Na verdade, o pneumococo é uma causa comum de otite média, pneumonia e meningite, sendo um agente principal de doenças graves em crianças jovens [Centers for Disease Control and Prevention2000]. É, por isso, especialmente importante estudar a dinâmica de transmissão de pneumococo entre as crianças que frequentam infantários, bem como tentar perceber de que forma esta dinâmica pode ser afectada por factores ambientais e características populacionais.

Existe uma grande variedade de estirpes de pneumococo, tendo sido descritos mais de noventa de cápsulas, ou serótipos, a que correspondem diferentes propriedades, tais como, um potencial patogénico variável [Brueggemann et al.2003]. Um estudo recente, usando testes *in vitro* sobre amostras clínicas, encontrou uma possível relação, com significado biológico, entre o tipo de cápsula e a capacidade de resistir ao sistema imunitário do hospedeiro [Weinberger et al.2009]. É possível que várias estirpes de pneumococo compitam na nasofaringe. Diversos tipos de competição foram estudados a nível teórico [Lipsitch1997, Zhang et al.2004] e os diferentes potenciais de colonização ou resistência a colonização foram testados para alguns serótipos em ratos [Lipsitch et al.2000]. Um número limitado de serótipos foi estudado e poucos deram diferenças estatisticamente significativas.

O primeiro tratamento utilizado contra pneumococo foi a soroterapia em estados iniciais de doença [Klugman2008]. A partir de uma amostra de expectoração, determinava-se o serótipo e injectava-se no paciente soro animal contendo anticorpos contra esse serótipo. Em 1940 esta terapia estava disponível para cinco serótipos. Entretanto, foram descobertos antibióticos eficazes, sendo de destacar a penicilina. Embora a mortalidade por doença pneumocócica nos muito jovens e idosos tenha continuado alta, a penicilina parecia uma droga milagrosa por

ser eficaz contra inúmeras doenças, algumas das quais quase sempre fatais, sem necessidade determinar o serótipo. Com o aumento do uso de penicilina no mundo, os tipos epidémicos mais invasivos, geralmente transmitidos por indivíduos doentes a pessoas saudáveis, mas que raramente se encontravam em indivíduos saudáveis, tornaram-se menos frequentes. Por outro lado, serótipos pediátricos menos virulentos, geralmente colonizadores por períodos prolongados e causa comum de otite média, começaram a substituir os tipos epidémicos nas infecções graves e a ganhar resistência aos antibióticos. Finalmente, na década de 80, uma vacina polissacárida eficaz contra as doenças pneumocócicas chegou ao mercado. Infelizmente, esta vacina não era eficiente em crianças. A primeira vacina conjugada a proteger contra a doença e o transporte de sete serótipos (PCV-7), também eficaz em crianças de menos de dois anos, só foi licenciada no ano 2000. Após o uso disseminado da vacina, a doença invasiva diminuiu, embora o transporte de pneumococo se tivesse mantido [Revai et al.2006]. A última vacina conjugada aprovada protege contra 13 serótipos (PCV-13).

O objectivo deste trabalho é estudar a dinâmica de transmissão de S. pneumoniae usando dados de um estudo longitudinal do estado de colonização, por esta bactéria, de crianças que frequentavam um infantário de Lisboa, Portugal [Sá-Leão et al.2008]. Estes dados foram obtidos em 1998, antes da introdução da vacina PCV7. Ao todo, 21 clones de pneumococo foram identificados correspondentes a 13 serótipos. Neste trabalho, o termo genótipo foi usado para designar o clone. A análise exploratória permitiu perceber que os dados eram insuficientes para modelar a transmissão para genótipos diferentes de forma diferencial. Considerou-se, por isso, que todos os genótipos tinham igual capacidade de ser transmitidos e de resistir à eliminação. Para caracterizar a dinâmica de transmissão de pneumococo no infantário, construiu-se um modelo conceitual da transmissão de pneumococo que pressupõe que todas as crianças amostradas estão em contacto entre si e se misturam, que a colonização de uma criança portadora de pneumococo pode ser diferente da colonização de uma criança não portadora e que as taxas de transmissão e de clearance se mantêm constantes. Embora tendo em conta o número de portadores de cada genótipo, considera-se que todos têm características iguais. Este modelo depende de quatro parâmetros: uma taxa de transmissão entre as crianças do infantário β, uma taxa de aquisição de pneumococo pela comunidade κ, um factor de competição entre genótipos ϕ e uma taxa de clearance μ. O valor destes parâmetros determina a dinâmica de transmissão própria ao infantário e foi estimado por inferência bayesiana. O modelo de verosimilhança

usado considera colonizações, super-colonizações, i.e. colonizações de crianças já portadoras, e clearances como processos Poisson de taxa constante entre os tempos de amostragem. Obtém-se o número observado de transições directamente dos estados amostrados, assumindo que as crianças não mudam de estado de colonização mais do que uma vez entre as amostras. As distribuições de probabilidade posterior conjunta dos parâmetros do modelo são estimadas por amostragem usando o método Monte Carlo com cadeias de Markov (MCMC).

Os intervalos de 90% de credibilidade e as médias, entre parênteses, obtidos foram [0.1248, 1.0625] (0.5974) para a taxa de transmissão genotípica β, [0.0014, 0.0217] (0.0107) para a taxa de aquisição genotípica na comunidade κ, [0.4348, 0.8670] (0.6280) para o parâmetro de competição ϕ e [0.2453, 0.3779] (0.3059) para a taxa de clearance genotípica μ.

Construiu-se um modelo estocástico de simulação baseado no mesmo modelo conceitual de transmissão para validar o modelo Poisson proposto para a estimação dos parâmetros e estudar o efeito de diferentes estratégias de amostragem. Este modelo foi implementado usando o método Gillespie, que tenta determinar o tempo e o tipo do próximo evento. Foram simuladas e amostradas histórias de colonização individuais, usando o estado de colonização das crianças amostrado na primeira visita ao infantário português acima referido, e os valores dos parâmetros estimados num estudo sobre transmissão pneumocócica em crianças de infantários finlandeses [Hoti et al.2009]. Dez intervalos de amostragem de um mês resultaram em estimativas enviesadas. Os parâmetros β, κ e μ apareceram sub-estimados e ϕ sobre-estimado. Muitos eventos de colonização e de clearances escaparam à observação. A análise dos dados simulados revela que os indivíduos efectuam, por vezes, transições muito próximas no tempo. O pressuposto de no máximo uma transição entre duas amostras para cada indivíduo não satisfaz quando o intervalo entre as amostras é de um mês. Alongando o período de estudo para trinta intervalos de amostragem mensais, aumentou a precisão da estimação, reduziu os intervalos de credibilidade, mas as estimativas mantiveram-se enviesadas. Diminuindo drasticamente o intervalo entre amostras, ou seja amostrando diariamente, quase todas as transições foram observadas e as probabilidades posteriores ficaram centradas em redor do valor correcto dos parâmetros, sendo esta, no entanto, uma estratégia de amostragem irrealizável.

Este trabalho demonstrou que não se pode assumir que as crianças só mudam de estado de colonização no máximo uma vez por mês, pelo menos considerando o valor dos parâmetros estimados para infantários finlandeses [Hoti et al.2009]. Propõe-se a inferência das histórias

individuais a partir dos estados observados como método alternativo que, por não depender deste pressuposto, permite estimar os parâmetros usando dados amostrados com um intervalo mais viável.

PALAVRAS-CHAVE Streptococcus pneumoniae, pneumococo, transmissão, inferência bayesiana, MCMC, Poisson

Contents

Chapter 1

Introduction

1.1 Motivation

The aim of this work is to characterize the transmission dynamics of *Streptococcus pneumoniae* in a Portuguese day-care center (DCC).

1.2 Literature Review

Pneumococcus epidemiology

Streptococcus pneumoniae, commonly referred to as pneumococcus, is a gram positive bacterium. Acquisition of pneumococcus always starts as colonization of the nasopharynx, the person becomes a carrier and is able to spread the bacteria to others. The close contacts between children attending DCCs, together with the immaturity of their immune system, make children younger than six years old a high risk group for nasopharyngeal carriage of pneumococcus [Bogaert et al.2004]. Carriage in older children and adults is less frequent. In cross-sectional studies of children attending Portuguese DCCs, pneumococcus was found in up to 71% of the children [Mato et al.2005]. Although colonization is generally asymptomatic, it can sometimes progress to infection. In fact, pneumococcus is a common cause of otitis media, pneumonia and meningitis, being a leading agent of serious illnesses in young children [Centers for Disease Control and Prevention2000].

Diversity

There is a wide variety of strains of pneumococcus. A common way to characterize a pneumococcal strain is by determining its type of capsule, or serotype. Over 90 different serotypes have been identified. There is some evidence that not all serotypes have the same pathogenic potential [Brueggemann et al.2003]. A recent study, using in vitro tests on clinical samples, found a possible biologically significant relationship between the type of capsule and the ability to resist the host immune system [Weinberger et al.2009]. Although, until recently, sampling methods usually only identified one of the clones from a sample [O'Brien and Nohynek2003], probably the dominant one, more than one clone of pneumococcus can be found in the same individual. It is possible that these pneumococcal clones compete with each other. Different types of competition were studied theoretically [Lipsitch1997] and some experiments were done in mice [Lipsitch et al.2000] but only some serotypes were studied and few resulted in statistically significant differences.

Antibiotics, vaccines and evolution of resistance

At the beginning of the 20^{th} century, there was a worldwide effort to develop a pneumococcal vaccine. Serum therapy had some success when applied to patients in early stages of illness. The serotype was determined from a sample of the patient's sputum and animal serum containing antibodies against that serotype was injected in the patient. This therapy was available for 5 serotypes by 1940. Around that time, effective antibiotics against pneumococcus were found, first sulfapyridine, for which resistant strains appeared within a couple of years, and then penicillin. Penicillin seemed effective in treating a wide range of diseases, some of which were almost always fatal like meningitis, without the need of serotyping. Although mortality rates for the very young and very old continued high, penicillin was thought as a miracle drug and efforts in vaccine research diminished. With the increased use of penicillin worldwide, the most invasive epidemic types, that usually spread from sick individuals to healthy contacts but were rarely carried by healthy individuals, became less frequent. On the other hand, less virulent pediatric serotypes, usually carried for prolonged periods and a common cause of otitis media, started to replace the epidemic types in serious infections and gain resistance to antibiotics. New efforts were made to develop a pneumococcal vaccine and a polysaccharide vaccine reached the market in the 1980s. However, it lacked efficacy in infants . The first conjugate vaccine that protected against disease and carriage of seven serotypes (PCV-7) was licensed only in the year 2000 (reviewed in Klugman, 2008). After widespread adoption of this vaccine, invasive disease has diminished but carriage of pneumococcus has remained the same [Revai et al.2006]. The last conjugate vaccine approved protects against 13 serotypes (PCV-13).

Modelling

A model is a simplified representation of reality. Since reality is too complex for us to understand, these simplifications can help us summarize what we know about a phenomenon, understand the importance of each part, and, if it is a good enough representation of reality, predict what would happen in different conditions, what would be the consequences of specific changes to the system.

Mathematical models are a quantitative description of a natural phenomenon. Reality is described through equations that incorporate all the parts of the system that we want to model.

There are many tools to analyze deterministic models, i.e. models that do not take chance into account. Events that happen at random, for example births and deaths, can be studied using deterministic equations as long as the population size is large. For small populations, chance plays an important role on the outcome. Stochastic models allow some randomness to be incorporated. Some stochastic models are event-driven, only integer number of persons are considered and probabilities of events are calculated from the rates [Keeling and Rohani2008]. One way to study stochastic models is to implement them computationally and analyze the results given the model and a chosen initial condition/population. Models described in equations depend on measurable quantities, for example the number of adults, but generally also depend on parameters. The value of parameters is not fixed and it affects the dynamic of the system. Using statistical inference, parameters values can be estimated so as to maximize the probability of the model reproducing some observed data [Keeling and Rohani2008, Gelman et al.1995].

1.3 Objectives

To create a model which relies on transmission parameters in order to characterize the pneumococcus transmission dynamics in the Portuguese DCC. To construct an estimation method to obtain confidence intervals for the parameter values. Finally, to evaluate the estimation method using simulated data for which the parameter values have been defined.

Chapter 2

Materials and Methods

2.1 The data

The data comes from a year-long longitudinal study in a day-care center (DCC) in Lisbon, Portugal, as described earlier [Sá-Leão et al.2008]. All the children attending three of the rooms of the DCC, 47 in total, participated in the study. They all played together in the playground. Samples were taken, starting in February 1998, on weeks 1, 5, 11, 15, 20, 29, 33, 38, 42, 47 and 50. The 9-week sampling interruption between weeks 20 and 29 was due to summer holidays. A pediatric nurse took the samples using alginate swabs which were later inoculated into selective culture medium for pneumococci and *Haemophilus influenzae*. Usually, a single pneumococcal strain was isolated. This strain was then serotyped, genotyped by pulsed-field gel electrophoresis (PFGE) and tested for antimicrobial susceptibility, defining the clones.

Overall, 414 samples were obtained, corresponding to 80% of the expected samples. Questionnaires were handed to the parents to fill in information on the children's illnesses and antimicrobial consumption in the month preceding each sampling. Throughout this dissertation, the term "genotype" will be used to refer to clones. Twenty-one pneumococcal genotypes were identified from thirteen distinct capsular serotypes and some untypeable (NT). The serotypes identified were, in decreasing order of abundance, 19F, 23F, 6B, 14, 10A, 19A, 9V, 11A, 16F, 18F, 15A, 8 and 23B. The two most abundant serotypes accounted for fifty percent of the isolates and the six most abundant serotypes accounted for more than eighty percent of the isolates.

Since this study was done independently from the original work, and the school has closed since, we were limited to the data originally collected.

2.2 Transmission model

The following model describes the transmission dynamics of pneumococcus between children in a DCC. Since the isolation technique generally only types one colony from the sample [Sá-Leão et al.2008, Ruoff et al.2003], children are considered to carry at most one genotype at any given time. Children are assigned a state corresponding to their carriage status. This state at any given time is zero if the child is not carrying pneumococcus. In a population where n_g genotypes are found, the state of a child carrying a given genotype is $g \in \{1, ..., n_g\}$. When the state of a child changes, a transition occurs. Children who are not carrying pneumococcus, i.e. non-carriers, can be colonized by a certain strain of pneumococcus if in contact with children carrying that

genotype. Let us refer to this type of transition as colonization. Children who are carrying a certain genotype of pneumococcus can also be colonized by pneumococcus of a different genotype. Given the isolation technique, for a child already colonized by a genotype to be found to carry another genotype, the new genotype has to become dominant in principle. In the study of infectious diseases, infection of an already infected host by a new strain that clears the one already present is called super-infection [May and Nowak1994]. Therefore, this type of transition will be referred to as super-colonization. Children who are carrying pneumococcus can also clear the bacteria and become non-carriers. This can happen for example by competition with another organism in the nasopharynx or by immune response. The transition of a carrier to a non-carrier will be called clearance.

A child already carrying a genotype is thought to be colonized at a rate that differs from non-carriers' colonization by a competition factor ϕ. If there is no competition between the genotypes, $\phi = 1$. If $\phi > 1$, it is easier for a genotype to colonize a host if this host was already carrying another genotype of the bacteria. If $\phi < 1$, there is competition between the genotypes, be it direct competition, for example consumption of the same resources or production of toxins against the other genotype, or indirect, for example cross-immunity, immunity to one genotype gives at least some immunity against the other genotype. Children are considered to clear colonization, i.e. become non-carriers, at a constant rate μ.

The rate $\lambda_c^{f,g}$ at which child c in state f changes to state $g \neq f$ is then

$$\lambda_c^{f,g} = \begin{cases} \phi\lambda^g & \text{if } f,g > 0 \\ \lambda^g & \text{if } f = 0 \\ \mu & \text{if } g = 0 \end{cases} \tag{2.1}$$

where λ^g is the baseline rate of colonization by genotype g.

Assuming that all the children mix equally and they all have contact with each other, the transmission rate within the group is proportional by a factor β to the fraction of contacts with a carrier. Defining C^g as the number of children in state g, the number of non-carriers will be referred as $S = C^0$ and the total number of carriers as $C = \sum_{g=1}^{n_g} C^g$. For $g \in \{1,...,n_g\}$, C^g is the number of carriers of genotype g. The within-group transmission rate of genotype g is then $\frac{\beta C^g}{n_c - 1}$. The rate at which the children acquire genotype g from contact with carriers from outside the group is the community rate of acquisition κ. The baseline rate of colonization by a genotype

g, i.e. the rate at which a non-carrier is colonized by genotype g, is then

$$\lambda^g = \frac{\beta C^g}{n_c - 1} + \kappa. \tag{2.2}$$

The parameters β, ϕ, κ and μ are assumed to be equal for all the children and all the genotypes. β, κ and μ are rates of unit $[\frac{1}{People \times Time}]$ and ϕ is adimensional.

Carriers are susceptible to clearances and to super-colonizations, but only by other genotypes than the one carried, and non-carriers are susceptible to colonizations. The rate of super-colonization, colonization and clearance for all the children is obtained by summing the rates of transitions across susceptible children,

$$\begin{aligned} \lambda^{scol} &= \sum_{f=1}^{n_g} C^f \left[\sum_{\substack{g=1,\\ g\neq f}}^{n_g} \phi \lambda^g \right] \\ \lambda^{col} &= S \sum_{g=1}^{n_g} \lambda^g \\ \lambda^{clear} &= \mu C. \end{aligned} \tag{2.3}$$

For the sake of clarity, from now on, the superscript ^{scol}, ^{col} or ^{clear} will be used when elements refer to, respectively, super-colonizations, colonizations or clearances.

2.3 Estimation Method

> The fundamental problem towards which the study of statistics is addressed is that of inference. Some data are observed and we wish to make statements, inferences, about one or more unknown features of the physical system which gave rise to this data. ... inference can most conveniently be thought of as concerned with statements about the unknown values of parameters.
>
> O'Hagan, 1994

In this work, statistical inference was used to get estimates for the parameters of the proposed transmission model using the observed states of the children at the sampling times. Let us define θ as the set of parameters to be estimated, $\theta = \{\beta, \phi, \kappa, \mu\}$ and let $\theta^* = \{\beta^*, \phi^*, \kappa^*, \mu^*\}$ be the value of θ that gave rise to the observed data. Since θ^* is not known, all that can be done is to try to know, given the prior knowledge $P(\theta)$ and observations O, which values of θ are most likely, i.e which values maximize the posterior probability distribution $P(\theta|O)$. Through the application of Bayes' theorem to statistical inference [Gelman et al.1995], it is known that

the posterior probability distribution $P(\theta|y)$ is directly proportional to the joint probability distribution $P(O,\theta)$,

$$P(\theta|O) \propto P(O,\theta). \tag{2.4}$$

and that the joint probability distribution can be calculated using the likelihood of the observations given θ, $L(O|\theta)$, and the prior distribution $P(\theta)$,

$$P(y,\theta) = P(y|\theta)P(\theta). \tag{2.5}$$

Given the proportionality, the values of θ that maximize the posterior probability distribution $P(\theta|O)$ also maximize the joint probability distribution $P(O,\theta)$.

Prior Probabilities

The prior probability distribution for ϕ is an Exponential with parameter $\frac{1}{ln(2)}$. This allows $P(\phi < 1) = P(\phi > 1)$ and $P(\phi <= 0) = 0$. Unless stated otherwise, β, κ and μ were given a non-informative prior. As a non-informative prior, a Uniform distribution was chosen which gives the same probability to all values between 0 and 100.

Likelihood

Using an individual-based stochastic adaptation of the transmission model, a likelihood function could be defined. The likelihood of transition times could be obtained from continuous-time survival analysis [Andersen et al.1997]. These transition times would need to be inferred from the state of the children at sampling times. This has been done, for example, to estimate similar parameters of pneumococcus transmission using a diferent set of data [Hoti et al.2009]. A different model was chosen that determines the likelihood of observed numbers of transitions per sampling interval.

By assuming that transitions are independent from each other and occur one-at-a-time, they can be modeled as a Poisson process. Consider a time interval $\tau_a = [a,b]$ of length $\Delta t = b - a$. Let N_{τ_a} be the counting process of the number events in the time interval τ_a as modeled by a Poisson process of constant rate λ. The probability of observing x events in this time interval depends only on λ and Δt:

$$P(N_{\tau_a} = x) = \frac{(\lambda\Delta t)^x e^{-\lambda\Delta t}}{x!}. \tag{2.6}$$

Equation (2.6) describes the Poisson probability distribution function.

If the Poisson process is inhomogeneous, i.e. its rate varies in time, λ in Equation (2.6) has to be replaced by $\int_a^b \lambda(t)dt$. To calculate the integral of a time varying transition rate adapted from Equation (2.3), we would need to know the number of non-carriers S and carriers C^g of each genotype g at every instant of the interval $[a,b]$.

In reality, however, these numbers are only known at the sampling times. In order to be able to calculate the likelihood, it was assumed that both the number of non-carriers and carriers of each of the genotypes remain constant between sampling times.

Let n_{st} be the number of sampling times T. For each sampling interval $\tau_j = [T_j, T_{j+1}]$ where $j \in \{1,...,n_{st}-1\}$ of length $\Delta_j = T_{j+1} - T_j$, given that Δ_j is about one month or less, it is assumed that children undergo at most one transition from time T_j to time T_{j+1}. Let s_t^c be the state of child c at time t. $\#\{c:\ s_{t_j}^c > 0\}$ refers to the number of children whose state at time T_j is greater than 0, meaning, in this case, the total number of carriers at time T_j. The number of observed transitions is then,

$$\begin{aligned} O_{\tau_j}^{scol} &= \#\{c:\ s_{T_j}^c > 0,\ s_{T_{j+1}}^c > 0,\ s_{T_j}^c \neq s_{T_{j+1}}^c\} \\ O_{\tau_j}^{col} &= \#\{c:\ s_{T_j}^c = 0,\ s_{T_{j+1}}^c > 0\} \\ O_{\tau_j}^{clear} &= \#\{c:\ s_{T_j}^c > 0,\ s_{T_{j+1}}^c = 0\} \end{aligned} \tag{2.7}$$

Equations (2.1) and (2.2) assume that all the children were present at all sampling times, and their state was known. There are, however, missing children in real sampled data, which means that, for a given sampling interval, some children were observed only at the beginning, and some only at the end. Only children that were observed both at the beginning and at the end of the sampling interval could account for transitions. The number of such children is $S_{obs}(T_j)$ and $C_{obs}^f(T_j)$ if, at the beginning of the sampling interval, they were non-carriers or carriers of genotype f, respectively. $C_{obs}(T_j) = \sum_{f=1}^{n_g} C_{obs}^f(T_j)$ is the total number of carriers at risk of a transition that this method can count. On the other hand, when assessing exposure to each serotype, $C^f(T_j)$ the total number of children carrying genotype f at T_j was used, independently of whether the child was or not sampled on T_{j+1}.

Summing the transition rates in Equation (2.1) by transition type, $\lambda^{scol}(\tau_j)$, $\lambda^{col}(\tau_j)$ and $\lambda^{clear}(\tau_j)$ could be obtained from Equations (2.12). Taking the above-mentioned considerations into account, for sampling interval τ_j, the rates of super-colonizations, colonizations and

clearances are, respectively,

$$\begin{aligned}
\lambda^{scol}(\tau_j) &= \sum_{f=1}^{n_g} C^f_{obs}(T_j)\left[\sum_{\substack{g=1,\\ g\neq f}}^{n_g} \phi\lambda^g(T_j)\right] \\
\lambda^{col}(\tau_j) &= S_{obs}(T_j)\sum_{g=1}^{n_g}\lambda^g(T_j) \\
\lambda^{clear}(\tau_j) &= \mu C_{obs}(T_j)
\end{aligned} \tag{2.8}$$

where

$$\lambda^g(T_j) = \frac{\beta C^g(T_j)}{n_c - 1} + \kappa\,. \tag{2.9}$$

Let us define the Poisson counting processes of the number of transitions in sampling interval τ_j, $N^{scol}_{\tau_j}$, $N^{col}_{\tau_j}$ and $N^{clear}_{\tau_j}$. The likelihood L of the observations is the probability that the counting processes $N^{scol}_{\tau_j}$, $N^{col}_{\tau_j}$ and $N^{clear}_{\tau_j}$ take the observed values $O^{scol}_{\tau_j}$, $O^{col}_{\tau_j}$ and $O^{clear}_{\tau_j}$ which can be calculated using Equation (2.6). Once the values of $S(T_j)$, $C^g(T_j)$ and the sampling interval length Δ_j are known, L depends only on the parameter values ($\theta = \{\beta, \phi, \kappa, \mu\}$) and the observed number of transitions. Let us consider the observed number of transitions for all sampling intervals by type, $O^{scol} = \{O^{scol}_{\tau_1}, ..., O^{scol}_{\tau_{n_{st}-1}}\}$, $O^{col} = \{O^{col}_{\tau_1}, ..., O^{col}_{\tau_{n_{st}-1}}\}$ and $O^{clear} = \{O^{clear}_{\tau_1}, ..., O^{clear}_{\tau_{n_{st}-1}}\}$. The likelihood of all the observations $O = \{O^{scol}, O^{col}, O^{clear}\}$ is

$$\begin{aligned}
L(O|\beta,\kappa,\phi,\mu) = \prod_{j=1}^{n_{st}-1} \; & P(\lambda^{scol}(\tau_j)\Delta_j = O^{scol}_{\tau_j}|\beta,\kappa,\phi) \\
& \times P(\lambda^{col}(\tau_j)\Delta_j = O^{col}_{\tau_j}|\beta,\kappa) \\
& \times P(\lambda^{clear}(\tau_j)\Delta_j = O^{clear}_{\tau_j}|\mu)
\end{aligned} \tag{2.10}$$

Markov Chain Monte Carlo

Having defined the prior probability distributions $P(\theta)$ and the likelihood $L(O|\theta)$, the joint distribution in Equation (2.5) could now be analyzed. Since θ is four-dimensional, it is difficult to calculate $P(\theta, O)$ analytically. A Markov Chain Monte Carlo (MCMC) was used to sample the prior distribution of each parameter then calculating the likelihood of the observations [Equation (2.10)] for these parameter values [Patil et al.2010]. An implementation of the DiffeRential Evolution Adaptive Metropolis (DREAM) method [Vrugt et al.2009], which allows faster convergence, was used that allows the creation of multiple parallel MCMCs [Salvatier2010]. For each estimation, 10 simultaneous MCMCs were used. Convergence was verified by Gelman-Rubin diagnostic [Gelman and Rubin1992], iterations stop when the values for R are sufficiently

close to one. Parameter estimates are given in terms of their posterior medians and 90% credibility intervals (90% CI), taken from the 5% to the 95% posterior quantiles.

2.4 Simulation Method

Stochastic Model

A population-based event-driven stochastic model was built based on the transmission model. It is a compartmental model with the population divided in $n_g + 1$ compartments, non-carriers and carriers of each one of n_g genotypes. Let $N_c^{f,g}(t)$ be the counting process that counts the number of transitions for child c from state f to state g up to time t, with $f, g \in \{0, 1, ..., n_g\}$ and $g \neq f$. For a study cohort of n_c children, the history of the $n_c \times (n_g + 1) \times n_g$ counting processes $N_c^{f,g}(t)$ at time t is denoted by H_t. The probability of transition is defined as

$$P(dN_c^{f,g}(t) = 1|H_t) = \lambda_c^{f,g}(t^-)I_c^f(t^-)dt$$

where $I_c^f(t^-)dt$ is an indicator function which takes value 1 if child c was in state f right before time t and 0 otherwise.

Since the the sum of counting processes is also a counting process [Andersen et al.1997], let us define $N^{scol} = \sum_{c=1}^{n_c} \sum_{f=1}^{n_g} \sum_{\substack{g=1,\\ g\neq f}}^{n_g} N_c^{f,g}$, $N^{col} = \sum_{c=1}^{n_c} \sum_{g=1}^{n_g} N_c^{0,g}$ and $N^{clear} = \sum_{c=1}^{n_c} \sum_{f=1}^{n_g} N_c^{f,0}$ as the counting processes for super-colonizations, colonizations and clearances, respectively. The probabilities of transition are then as follows:

$$\begin{cases} P(dN^{scol}(t) = 1|H_t) = \lambda^{scol}(t^-)dt \\ P(dN^{col}(t) = 1|H_t) = \lambda^{col}(t^-)dt \\ P(dN^{clear}(t) = 1|H_t) = \lambda^{clear}(t^-)dt \end{cases} \tag{2.11}$$

where

$$\begin{aligned} \lambda^{scol}(t) &= \sum_{f=1}^{n_g} C^f(t)[\sum_{\substack{g=1,\\ g\neq f}}^{n_g} \phi\lambda^g(t)] \\ \lambda^{col}(t) &= S(t)\sum_{g=1}^{n_g} \lambda^g(t) \\ \lambda^{clear}(t) &= \mu C(t) \end{aligned} \quad . \tag{2.12}$$

$S(t)$ is the number of non-carriers at time t, $C^f(t)$ is the number of carriers of genotype f at time t and $C(t) = \sum_{f=1}^{n_g} C^f(t)$ is the total number of carriers at time t. Similar to Equation (2.2),

$$\lambda^g(t) = \frac{\beta C^g(t)}{n_c - 1} + \kappa. \tag{2.13}$$

Computational Implementation

The Gillespie algorythm [Gillespie1977], is a stochastic simulation algorythm. The next event is chosen from the rates defined by the model and a random number, the higher the rate, the higher the probability of occuring. An implementation of the Gillespie algorithm [Coelho2009] was used to simulate data from the stochastic model described. A length, in months, of the simulation was chosen. The starting number of individuals in each compartment was taken from the first sample of the portuguese DCC. The values for the the transmission parameter β, the community rate of acquisition κ, the competition parameter ϕ and the clearance rate μ were set as the values estimated in a study of Finnish DCCs using a similar transmission model [Sá-Leão et al.2008]. The times of transitions were returned. The results of this compartmental simulation were used to simulate a possible pneumococcal acquisition history for each individual. As shown in the pseudo code below, for each transition returned by the Gillespie simulation, one of the susceptible individuals was chosen and this transition was added to the individual's history.

```
FUNCTION dataMatrix(simulationResults)
   SELECT individuals, transitions, initialPopulation FROM simulationResults
   FOR each individual in individuals
       SET the current state as the state of the individual in the initialPopulation
       CREATE a list of the individual's states
       CREATE a list of the individual's transition times
   ORDER transitions by time of transition
   FOR each transition in transitions
       CHOOSE one individual from the individuals with state equal to the initial state of the transition
       SET the current state of the individual as the final state of the transition
       ADD final state of the transition to the child's states
       ADD time of event to the individual's transition times
   RETURN states and transition times for all individuals
```

From the resulting histories, a table was created with the state of each individual at the sampling times. This table is similar to the one obtained from the data sampled from the DCC and it was analyzed the same way. To estimate the parameters, only data starting at month 11 of the simulation was used, to give time for the system to approach its equilibrium.

Chapter 3

Results

3.1 Exploratory Data Analysis

The preliminary analysis of the Portuguese DCC data is presented here. The number of children in each of the states at the sampling times can be seen in Figure 3.1. The genotype 19F-Pn3 was dominant, decreasing just in the last samples. Only 7, of the 21 genotypes sampled, were found at the first sampling time (19F-Pn3, 23F-Pn1, 9V-Pn2, NT-Pn5, 11A-Pn8, 6B-Pn12, NT-Pn17). The other genotypes appeared throughout the year. This suggests that interactions with people outside the study cohort was important for the observed transmission dynamics. Some genotypes colonized only one child, then disappeared (NT-Pn17, NT-Pn18, NT-Pn10, 8-Pn15, 23B-Pn16). The average number of non-carriers and carriers was 12.8 and 18.3, corresponding to 59.2% and 40.8% of sampled children, respectively.

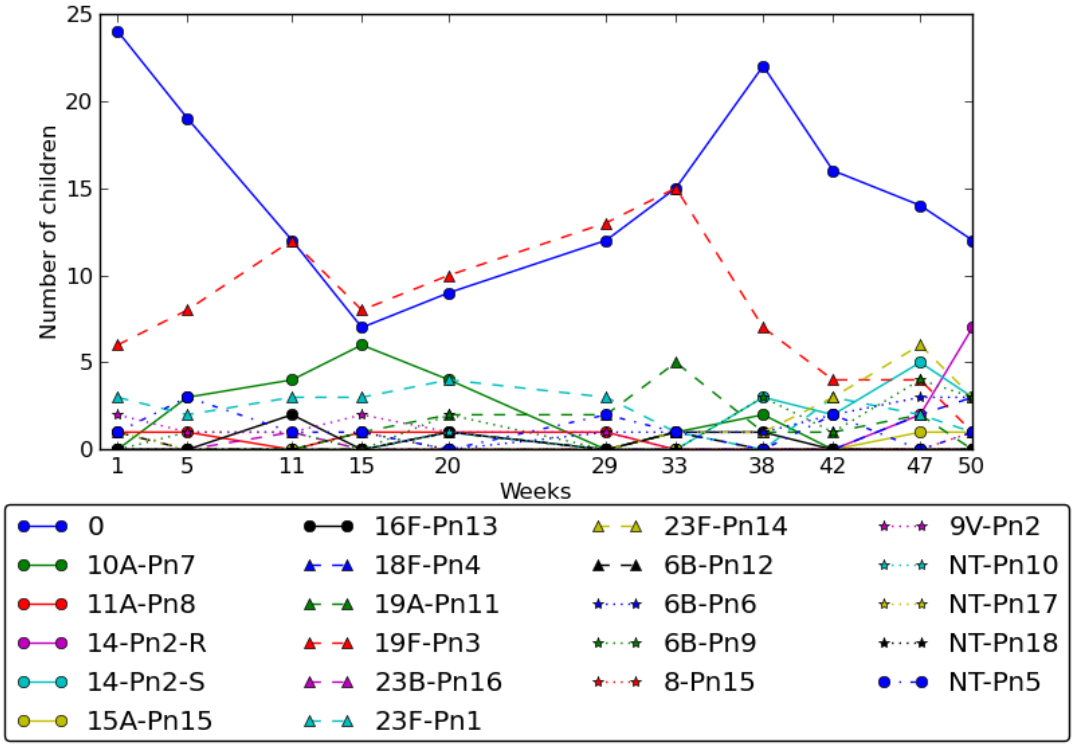


Figure 3.1: **Samples.** Number of non-carriers (0) and carriers of each genotype at the sampling times.

The number of times children were sampled at two consecutive sampling times and the observed states is presented in 3.1. A total of 280 pairs of observations were obtained. The numbers on the diagonal, in bold, are the number of times the state of a child remained the same at two consecutive sampling times. All the other numbers represent the number of

transitions observed, with the starting state on the row label and the end state on the column label, assuming the child did not change state more than once between these sampling times. Under this assumption, 164 transitions were observed, from which 61 were colonizations, 55 were clearances and 48 were super-colonizations. Since 21 genotypes were found, 21 colonizations, 21 clearances and 420 super-colonizations were possible. Many did not occur in this dataset. Even though there might be genotype-specific transmission dynamics, there is not enough data to discriminate it. For this reason, the transmission model considers all genotypes equal in their ability to colonize, to withstand colonization and to withstand clearance.

Table 3.1: **Transition Matrix.** Number of children observed in the state on the row label at one sampling time and in the state on the column label at the next sampling time. Non-zero numbers of transitions are in bold orange. In the row labels, the genotype name is presented along with the corresponding state index, in parentheses, and 0 stands for non-carrier. In the column labels, only the state index is used.

state at T_j \ state at T_{j+1}	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Total
0 (0)	**54**	**5**	0	**2**	**3**	**1**	**2**	**2**	**3**	**20**	**1**	**4**	**8**	0	**3**	**4**	0	**1**	0	0	**1**	**1**	115
10A-Pn7 (1)	**4**	**2**	0	0	0	0	**1**	0	**1**	**3**	0	0	0	0	**1**	0	0	0	**1**	0	0	0	13
11A-Pn8 (2)	**2**	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
14-Pn2-R (3)	0	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
14-Pn2-S (4)	**2**	0	0	0	**4**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**	8
15A-Pn15 (5)	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
16F-Pn13 (6)	**1**	0	0	0	**1**	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
18F-Pn4 (7)	**1**	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
19A-Pn11 (8)	**5**	0	0	0	0	0	**1**	**1**	**3**	0	0	0	0	0	0	**1**	0	0	0	0	0	0	11
19F-Pn3 (9)	**21**	**3**	**1**	**1**	**1**	0	0	0	**1**	**25**	0	**1**	0	**1**	0	**2**	0	**1**	0	0	0	**2**	60
23B-Pn16 (10)	**1**	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	1
23F-Pn1 (11)	**6**	0	0	**1**	0	0	0	**1**	0	0	0	**8**	0	0	0	0	0	0	0	0	0	0	16
23F-Pn14 (12)	**5**	0	0	**3**	0	0	0	0	0	0	0	0	**1**	0	0	0	**1**	0	0	0	0	0	10
6B-Pn12 (13)	**1**	**1**	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	2
6B-Pn6 (14)	**1**	0	0	0	0	0	0	0	0	0	0	**1**	0	0	**6**	0	0	0	0	0	0	0	8
6B-Pn9 (15)	**3**	0	0	0	**2**	0	0	0	0	0	0	0	0	0	0	**4**	0	0	0	0	0	**1**	10
8-Pn15 (16)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0
9V-Pn2 (17)	0	**1**	0	0	0	0	0	0	0	**2**	0	0	0	0	0	0	0	**3**	0	0	0	0	6
NT-Pn10 (18)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0
NT-Pn17 (19)	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	**0**	0	0	1
NT-Pn18 (20)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	**0**	0	1
NT-Pn5 (21)	**2**	**1**	0	0	**1**	0	**1**	0	0	0	0	0	0	**1**	0	**1**	0	0	0	0	0	**2**	9
Total	109	13	2	9	12	2	5	4	8	50	1	15	9	2	10	13	1	5	1	0	1	8	280

3.2 Parameter Estimation

The estimates of the parameters using the Portuguese DCC data are here presented. To estimate the parameters, the number of people at risk of a transition, i.e. observed at two consecutive sampling times, is assumed to remain constant, as shown in Figure 3.2. The sampling interval from week 20 to 29 is not used. A total of 10 sampling times, 9 sampling intervals, was considered. The level of exposure to each genotype is also assumed to stay constant between sampling times, so the number of carriers of each genotype is assumed to stay equal to what was observed at the beginning of the sampling interval.

The results of the estimation are shown in Figure 3.3. The posterior median for the genotype transmission parameter β was 0.6136 per child per month (90% CI [0.1248, 1.0625]) (Table 3.3(c)), and the posterior mean was 0.5974. This is a little bit higher than but close to the serotype transmission parameter mean estimated for Finnish DCCs, 0.53 per child per month [Hoti et al.2009]. The 90% CI is very large, encompassing the Finnish value. The community rate of acquisition κ was estimated to be 0.0104 per month per genotype in a non-carrying child (90% CI [0.0014, 0.0217]) (Table 3.3(c)), and the posterior mean was 0.0107. This estimate is much higher than the Finnish posterior mean, 0.0059 [Hoti et al.2009]. For this parameter, the Finnish posterior mean also falls within the 90% CI obtained in the present study. All influences outside from the sampled children affect transmission through this parameter. This is therefore expected, since only three rooms of the Portuguese DCC were sampled and the study, unlike the Finnish study, did not discriminate the family rate of pneumococcal acquisition. The posterior median for the competition parameter ϕ was 0.6123 (90% CI [0.4348, 0.8670]) (Table 3.3(c)), and the posterior mean was estimated to be 0.6280. Since 1 is not included in the credibility interval, there seems to be competition between the genotypes, being a carrier seems to offer some resistance to being colonized by other genotypes. The study of the Finnish DCC data reached the same conclusion at the serotype level, estimating a posterior mean of 0.68 [Hoti et al.2009]. For this parameter, the Finnish posterior mean also falls within the 90% CI obtained

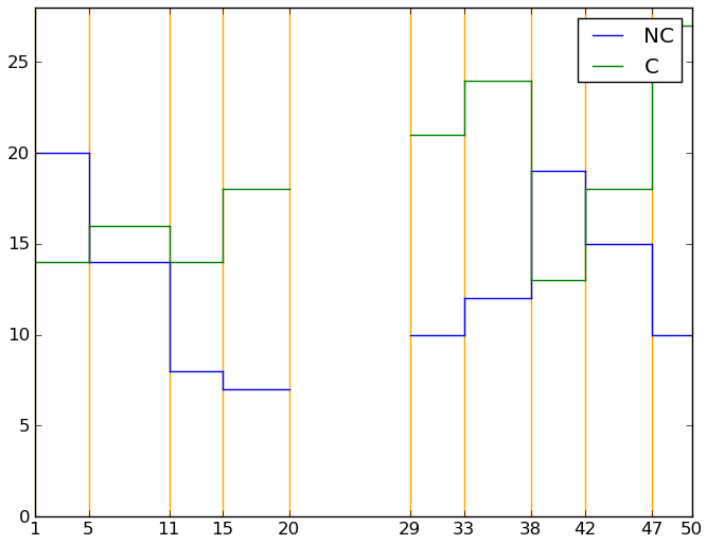
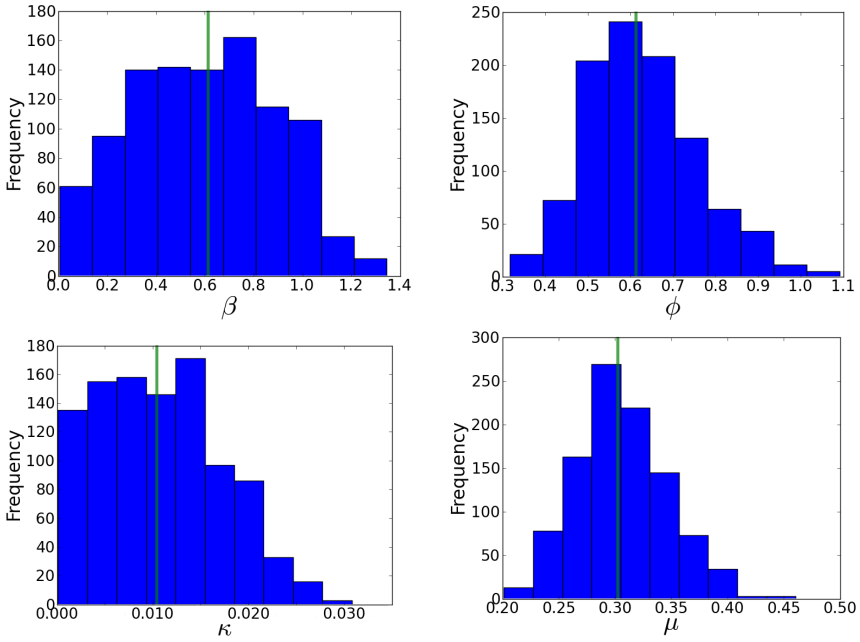
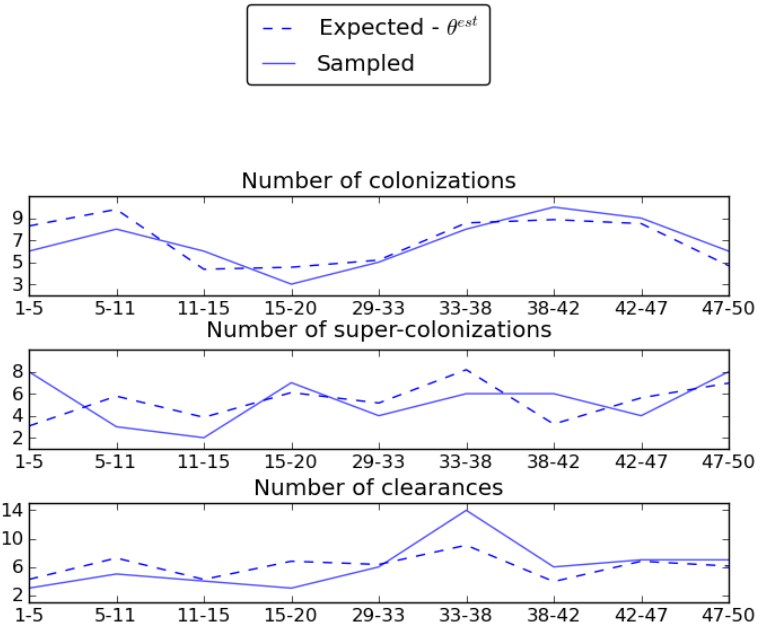


Figure 3.2: **Sampling result.** Number of non-carriers (S) and carriers (C) at risk of transition at sampling times, in weeks, from the DCC samples.



(a) Posterior distributions



(b) Number of transitions per sampling interval

	Median	90% CI		Range	Priors
β	0.6136	0.1248	1.0625	0.9377	U(0,100)
ϕ	0.6123	0.4348	0.8670	0.4323	E($\frac{1}{ln(2)}$)
κ	0.0104	0.0014	0.0217	0.0203	U(0,100)
μ	0.3023	0.2453	0.3779	0.1325	U(0,100)

(c) Numerical results

Figure 3.3: **Estimation results.** Using DCC data. a) The green line is the median. b) θ^{est} is the estimated median. c) Range is the range of the 90% CI, U(0,100) means a Uniform distribution from 0 to 100, E($\frac{1}{ln(2)}$) means an Exponential distribution of parameter $\frac{1}{ln(2)}$.

in the present study. The clearance rate μ was estimated to be 0.3023 per child per month (90% CI [0.2453, 0.3779]) (Table 3.3(c)), with a posterior mean of 0.3059. This is significantly lower than the rate estimated for the Finnish DCC, 0.69 [Hoti et al.2009]. This would suggest that children in the Portuguese DCC were less able to clear colonization than children in Finnish DCCs.

As can be seen in Figure 3.3(b), the average expected number of transitions, setting the parameters's values with the estimated medians and considering the number of carriers and non-carriers at each sampling time, follows quite closely the number of sampled transitions. This seems to confirm that the estimation method found values for the parameters that best "fit" the sampled number of transitions. The bias comes from the distance between the sampled number of transitions and the number of transitions that occured in the simulation.

The precision of the estimation can be measured from the range of the 90% CI (credibility interval), the larger the range, the lower the precision. To compare precision between parameters the ratio between the range of the 90% CI and the estimated median was used. This ratio was 1.5282 for β, 1.9496 for κ, 0.7060 for ϕ and 0.4384 for μ. This means precision was much lower for β and κ.

A regression analysis of the posterior MCMC samples for the parameters was done to check for correlations (Figure 3.4). The correlation between β and κ was found to be strong but other parameters did not appear to be highly correlated. The fact that the posterior samples for β and κ were so strongly correlated explains the lack of precision for these parameters. Given this correlation, a higher value for β and a lower valuer for κ was as likely, for the model, as the opposite, causing more values to be accepted for both of these parameters.

3.3 Validation

The results of estimating the parameters using data simulated with chosen parameters values are presented here. Figure 3.5 shows the compartmental result of running a 40-month long simulation starting with the 39 states that were sampled at the first visit to the Portuguese DCC. At any moment, the sum of the number of non-carriers (S) and carriers (C) is the number of individuals, in this case 39. To allow for the system to reach its equilibrium, the first ten months of the simulation were not used.

The parameters were estimated using 11 monthly samples from this simulation, 10 sampling

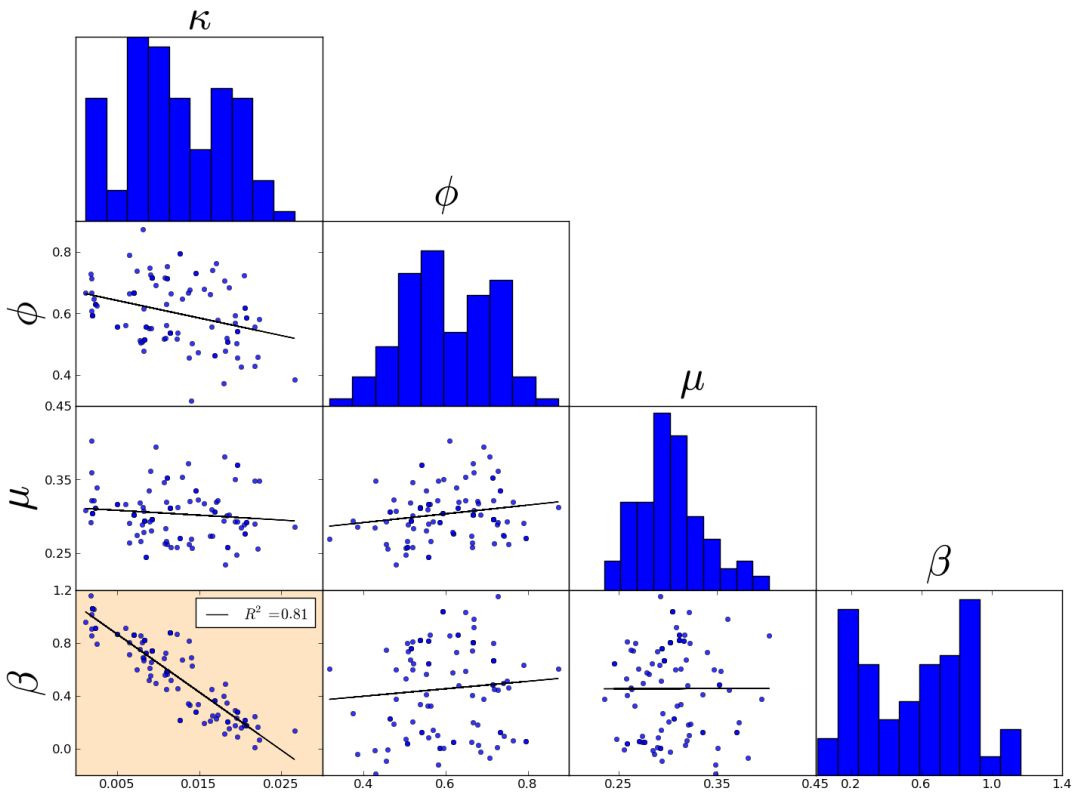


Figure 3.4: **Regression Analysis.** Correlation between MCMC samples for the parameters using DCC data measured through linear regression. Only the R-square of correlations of p-value less than 0.001 are shown.

intervals, and making the same assumptions as before. The number of individuals at risk for a transition used for the estimation are shown in Figure 3.6. The results of this estimation were biased [Figure 3.7]. The values of ϕ and μ chosen for the simulation were not included in the 90% CI [Figure 3.7(c)]. β, κ, and μ were underestimated and ϕ was over-estimated.

As a measure of accuracy, the difference between the chosen values θ^* and the estimated median θ^{est} was used. The closer this value to 0, the more accurate the estimation. This difference was 0.256 for β, 0.0008 for κ, -0.4018 for ϕ and 0.295 for μ.

The individual histories simulated for this period are shown in Figure 3.8. As can be seen in this figure, sometimes an individual undergoes two transitions in very short periods of time. The assumption that the individuals undergo at most one transition between two sampling times is not satisfied. Some transitions are hidden by a second transition in the same sampling interval and therefore the observed number of transitions is lower than simulated, see Figure 3.7(b). μ is estimated from the number of observed clearances, β and κ from the number of observed colonizations and super-colonizations and ϕ from the number of observed super-colonizations.

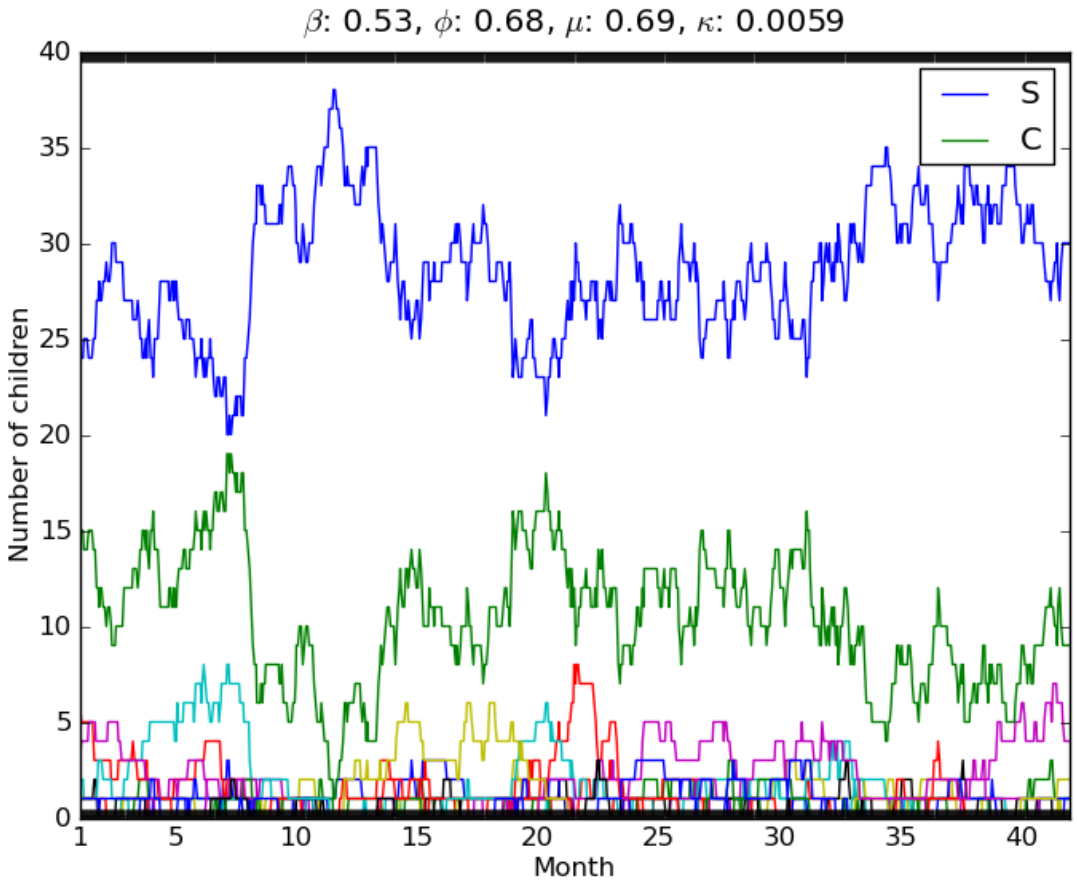


Figure 3.5: **Simulation Result.** Number of non-carriers (S), total number of carriers (C) and carriers of each genotype (not labeled lines). The values above are the values used for the simulation (θ^*)

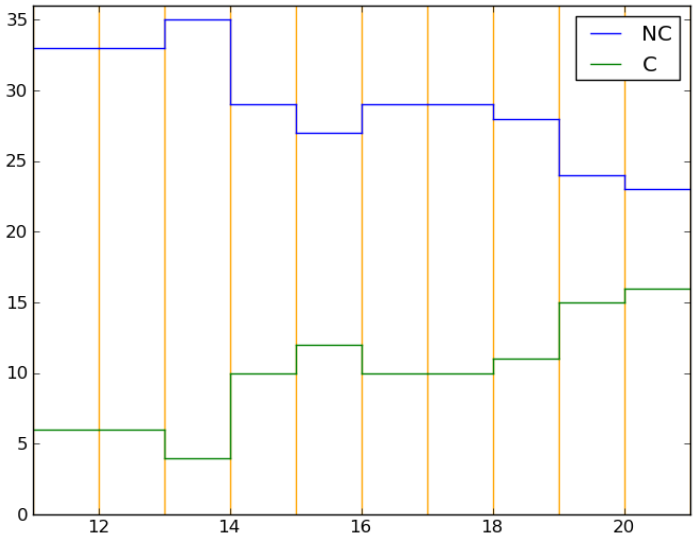
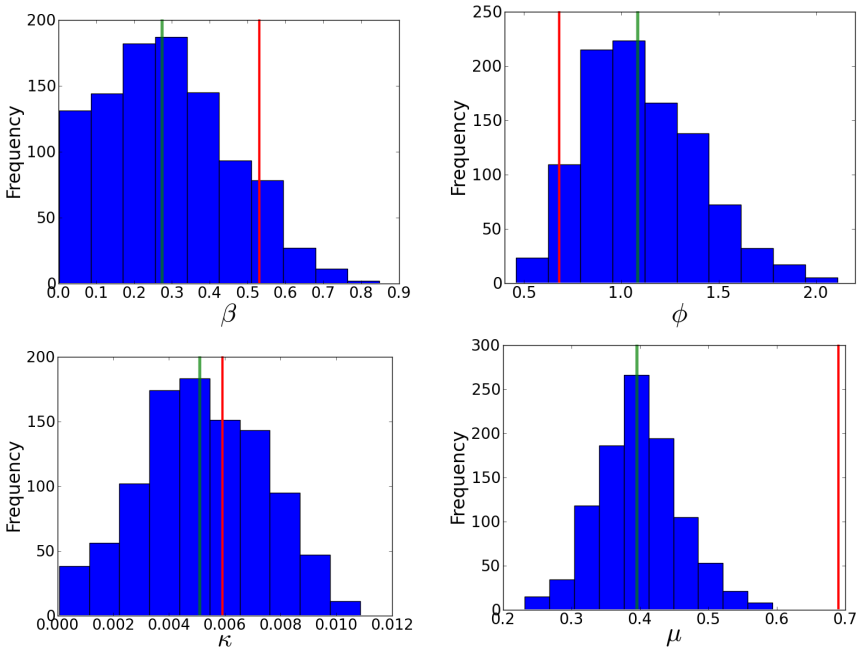
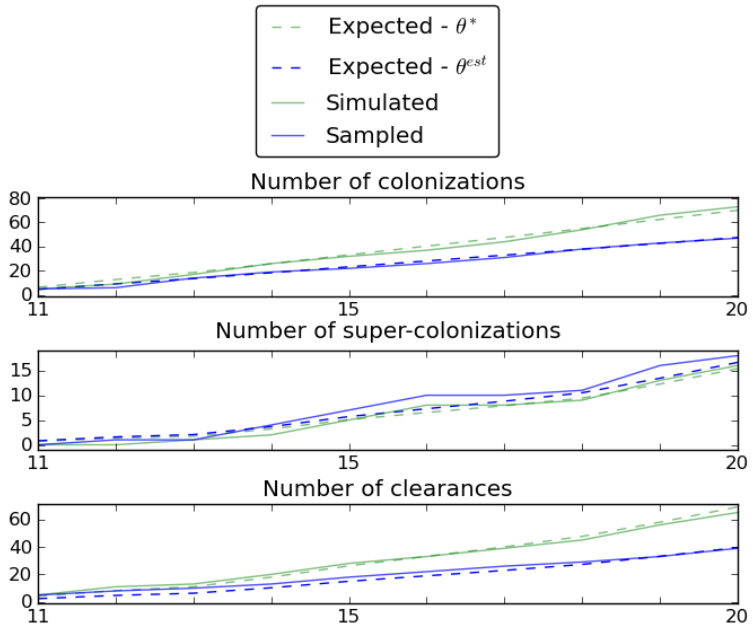


Figure 3.6: **Sampling result.** Number of non-carriers (S) and carriers (C) at risk of transition at sampling times, in months, from month 11 to month 21 of a simulation with 39 individuals.



(a) Posterior distributions



(b) Cumulative number of transitions per sampling interval

	Median	90% CI		Range	Priors	θ^*
β	0.2740	0.0364	0.5817	0.5453	U(0,100)	0.5300
ϕ	1.0818	0.6993	1.6336	0.9344	E($\frac{1}{ln(2)}$)	0.6800
κ	0.0051	0.0015	0.0087	0.0072	U(0,100)	0.0059
μ	0.3950	0.3055	0.5045	0.1990	U(0,100)	0.6900

(c) Numerical results

Figure 3.7: **Estimation results.** Using 10 months of simulated data with 39 individuals sampled once a month. a) The green line is the median, red line is the value that was defined for the simulation. b) θ^{est} is the estimated median. c) Range is the range of the 90% CI, U(0,100) means a Uniform distribution from 0 to 100, E($\frac{1}{ln(2)}$) means an Exponential distribution of parameter $\frac{1}{ln(2)}$.

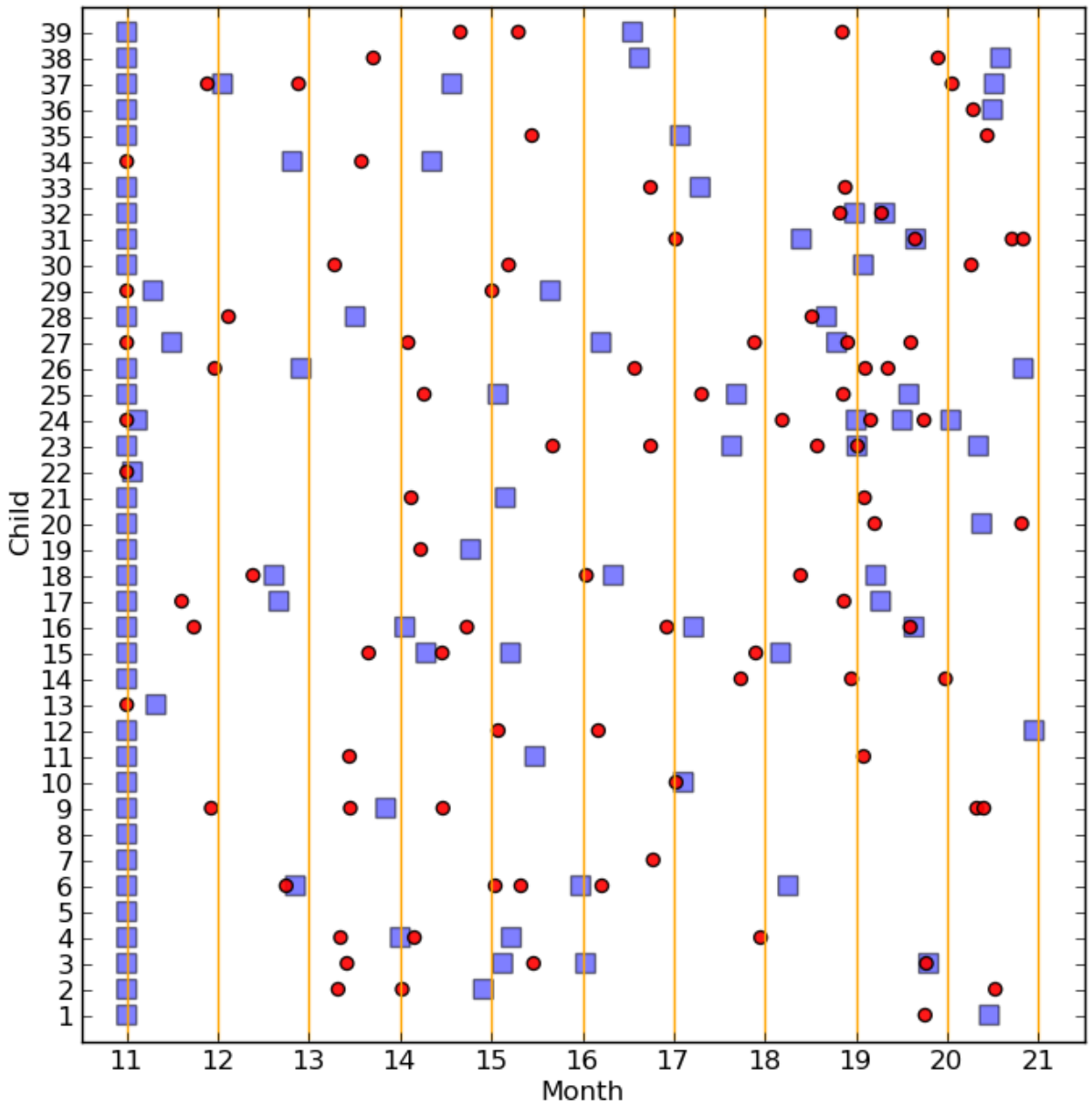


Figure 3.8: **Individual carriage histories.** Initial state of the individuals (at the first time) and states after transitions. The blue square means non-carrier and the red circle means carrier.

Since colonizations and clearances were much more frequent, β, κ and μ were the most affected by the bias in the observed number of transitions, causing them to be under-estimated. On the other hand, since super-colonizations were much less frequent, the observed super-colonizations were proportionally much closer to the simulated numbers. Thus, the over-estimation of ϕ is a compensation for the under-estimation of β and κ.

The ratio between the range of the 90% CI and the estimated median was 1.9902 for β, 1.4103 for κ, 0.8637 for ϕ and 0.5038 for μ, which means that β and κ were estimated with the least precision. As when using the DCC data, these two parameters appeared highly correlated

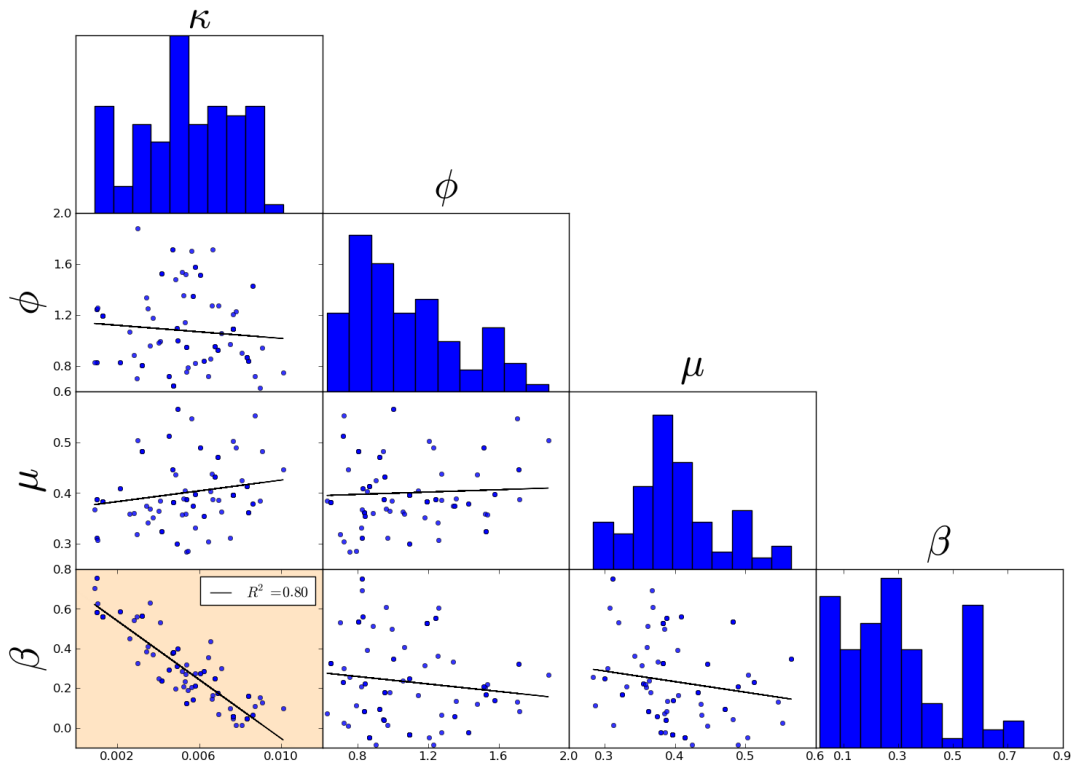


Figure 3.9: **Regression Analysis.** Correlation between MCMC samples for the parameters using data from month 11 to month 21 of a simulation with 39 individuals measured through linear regression. Only the R-square of correlations of p-value less than 0.001 are shown.

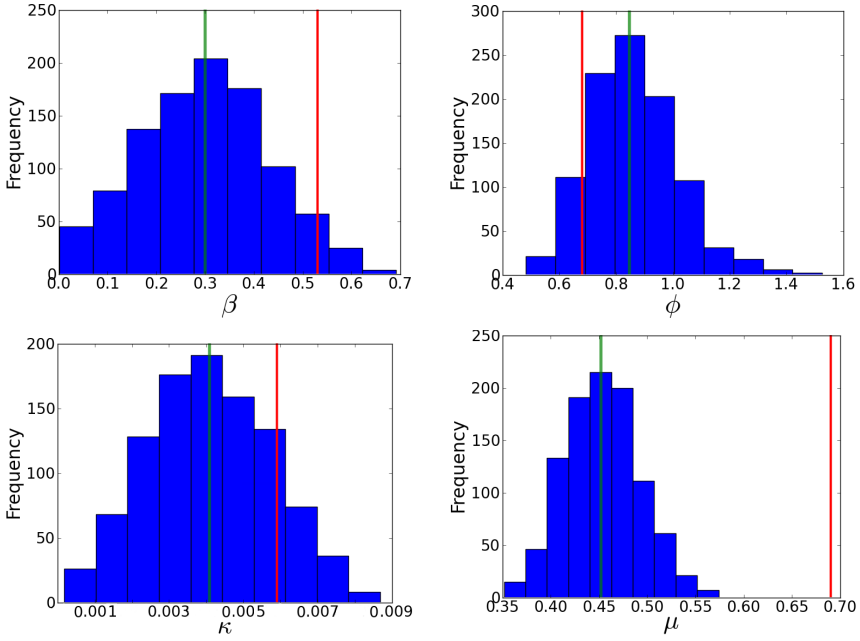
[Figure 3.9].

Longer Study Period

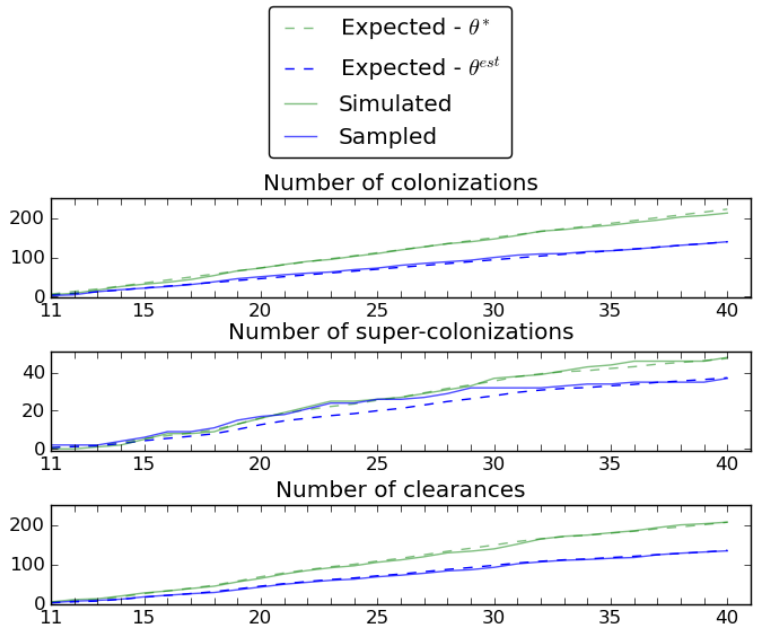
The effects of sampling monthly for a longer period are presented here. Figure 3.10 shows the estimation results from sampling monthly for thirty-one months using the same simulated data and thirty sampling intervals. The same bias is apparent as before, β, κ and μ were under-estimated and ϕ was over-estimated. The values of β and μ chosen for the simulation were not included in the 90% CI [Figure 3.10(c)].

Compared to the results of using ten monthly sampling intervals, the estimated medians for β, ϕ and μ were a little bit closer to the chosen values [Figure 3.7(c) and 3.10(c)], the difference between the parameter's value and the estimated median was 0.2301 for β, 0.0018 for κ, -0.1665 for ϕ and 0.238 for μ, so accuracy was slightly improved for all parameters except κ.

The ranges of the 90% CI were smaller for all parameters [Figure 3.7(c) and 3.10(c)], corresponding to an 18% decrease for β, 48% decrease for ϕ, 25% decrease for κ and 38% decrease for μ. A longer study period seems to improve precision for all parameters.



(a) Posterior distributions



(b) Cumulative number of transitions per sampling interval

	Median	90% CI		Range	Priors	θ^*
β	0.2999	0.0759	0.5216	0.4457	U(0,100)	0.5300
ϕ	0.8465	0.6381	1.1178	0.4798	E($\frac{1}{ln(2)}$)	0.6800
κ	0.0041	0.0015	0.0069	0.0054	U(0,100)	0.0059
μ	0.4520	0.3917	0.5158	0.1242	U(0,100)	0.6900

(c) Numerical results

Figure 3.10: **Estimation results.** Using 30 months of simulated data with 39 individuals sampled once a month. a) The green line is the median, red line is the value that was defined for the simulation. b) θ^{est} is the estimated median. c) Range is the range of the 90% CI, U(0,100) means a Uniform distribution from 0 to 100, E($\frac{1}{ln(2)}$) means an Exponential distribution of parameter $\frac{1}{ln(2)}$.

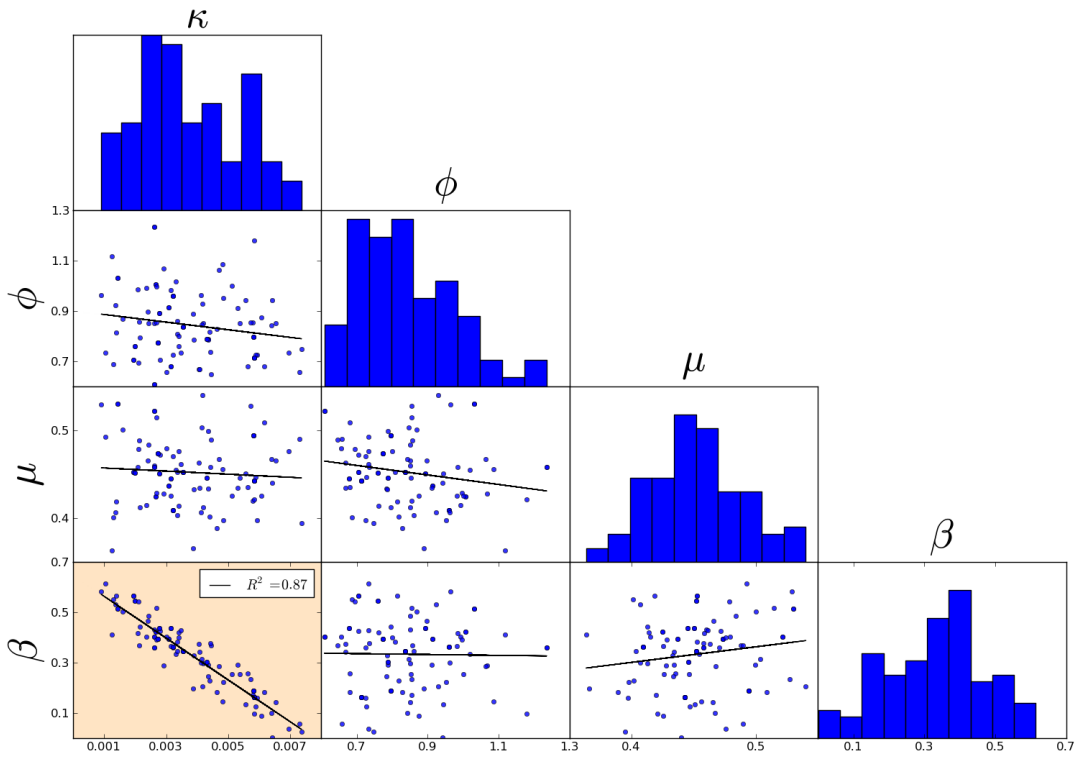


Figure 3.11: **Regression Analysis.** Correlation between MCMC samples for the parameters using data from month 11 to month 41 of a simulation with 39 individuals measured through linear regression. Only the R-square of correlations of p-value less than 0.001 are shown.

The correlation between β and κ with this data is even more obvious [Figure 3.11].

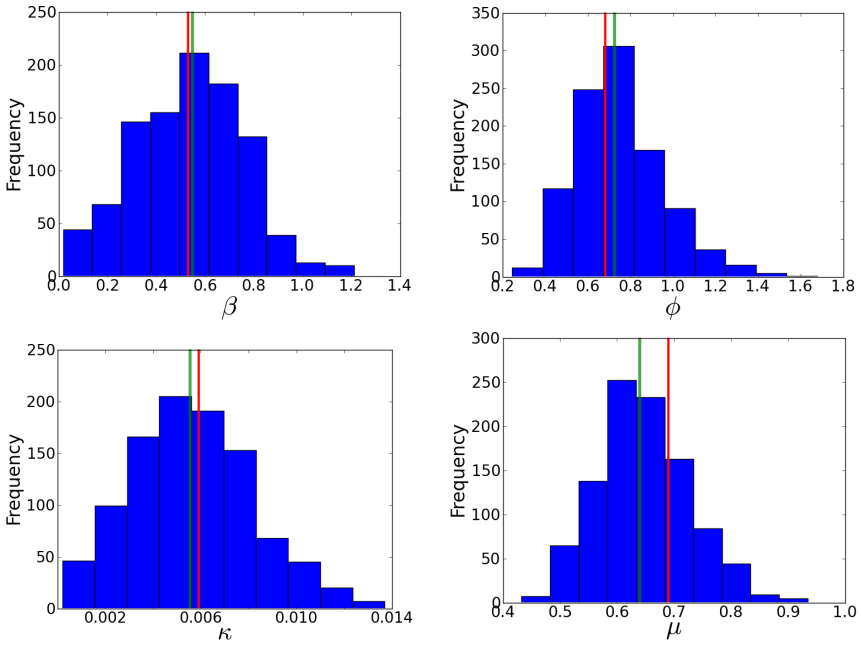
Shorter Sampling Interval

The results of shortening the sampling interval are shown here. Since time between consecutive events for the same individual appeared to be very small sometimes [Figure 3.8], a sampling interval of one day was chosen to allow for the assumption of at most one transition per individual per sampling time to be satisfied.

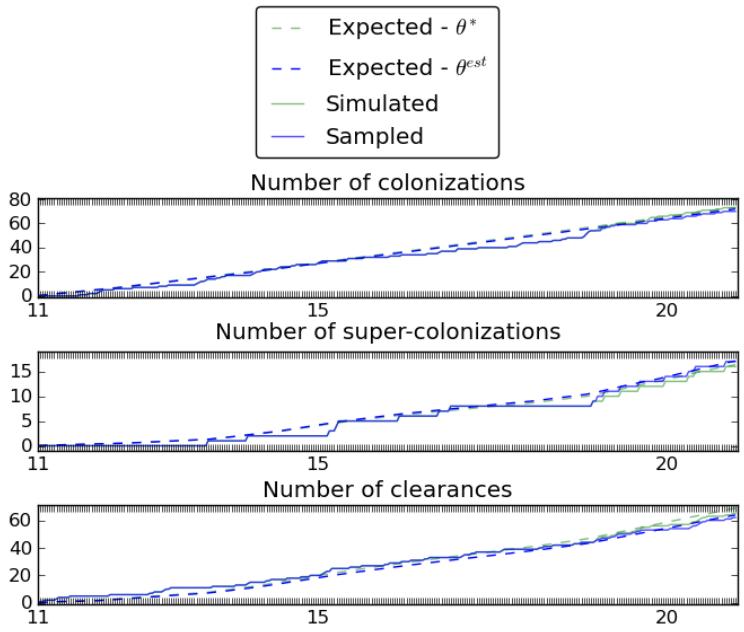
The results of sampling daily for 10 months are shown in Figure 3.12. With very few exceptions, the number of sampled transitions corresponded to the number of transitions simulated. No significant sampling bias was introduced with this sampling interval [Figure 3.12(b)].

Comparing to the results of taking 10 monthly sampling intervals, the range of the 95% CI was smaller, precision improved, only for ϕ. For all other parameters, precision was lower.

The difference between the parameter values and the estimated median was -0.0165 for β, 0.0004 for κ, -0.0434 for ϕ and 0.0503 for μ, all of which are much closer to 0. Such a short sampling interval improves accuracy immensely. However, this is not pracical and the use of an



(a) Posterior distributions



(b) Cumulative number of transitions per sampling interval

	Median	90% CI		Range	Priors	θ^*
β	0.5465	0.1463	0.8700	0.7237	U(0,100)	0.5300
ϕ	0.7234	0.4615	1.1204	0.6589	E($\frac{1}{ln(2)}$)	0.6800
κ	0.0055	0.0017	0.0102	0.0085	U(0,100)	0.0059
μ	0.6397	0.5189	0.7892	0.2703	U(0,100)	0.6900

(c) Numerical results

Figure 3.12: **Estimation results.** Using 10 months of simulated data with 39 individuals sampled thirty times a month. a) The green line is the median, red line is the value that was defined for the simulation. b) θ^{est} is the estimated median. c) Range is the range of the 90% CI, U(0,100) means a Uniform distribution from 0 to 100, E($\frac{1}{ln(2)}$) means an Exponential distribution of parameter $\frac{1}{ln(2)}$.

alternative method is mentioned in the discussion.

Chapter 4

Discussion

Longitudinal data on pneumococcal carriage in a cohort of Portuguese day care children was analysed. Rates of pneumococcal acquisition, conditional on genotype specific exposure, were estimated within a Bayesian framework. To use the states of the children at sampling times to count the number and types of transitions that occured in the sampling interval, the assumption was made that the children did not undergo more than one transition between two sampling times.

The parameters were estimated using the DCC data. The posterior mean of the genotype transmission parameter β was estimated to be 0.5974 per child per month, that of the community rate of acquisition κ was estimated to be 0.0107 per month per genotype for non-carrying children, that of the competition parameter ϕ was estimated to be 0.6280 and that of the clearance rate μ was estimated to be 0.3059 per child per month.

The method was subsequently applied to a set of simulated data showing that the assumption of no more than one transition per sampling interval was not appropriate when the sampling interval is as long as one month. β, κ are μ were under-estimated while ϕ was over-estimated. The bias observed when using a monthly sampling interval on simulated data presumably also affected the estimation made from the Portuguese DCC data. Since the estimated values for β and κ were already higher than those estimated for Finnish day-care children [Hoti et al.2009], if β and κ were in fact under-estimated, their true value would be even higher, which can possibly mean that the transmission dynamic of pneumococcus between day-care children in Portugal is faster than in Finland. μ was estimated to be much lower than was estimated for Finnish DCC children [Hoti et al.2009]. If it was under-estimated, the true average clearance rate might be closer to the Finnish estimate. The value estimated for ϕ was slightly lower than the one estimated from the Finnish dataset. If ϕ was over-estimated, then the true value would be even lower, in which case competition between genotypes would be even more evident.

The precision of the estimation of β and κ was very low. A high correlation was found in the samples taken for these two parameters, high values for one of them correlated with low values for the other. This resulted in very wide posterior distributions and credibility intervals, giving little confidence in the estimates. The precision was found to be improved when using a longer study period (30 months).

A real difference between capacities to clear pneumococcus in these two countries might not be realistic, but the difference in the estimated clearance rates could be an artifact. The

model used does not consider that an individual carrying a genotype (strain) can be colonized by the same genotype (strain). These types of transitions would not be visible from the data directly, but if they occur and the model does not consider them, the clearance rate will appear lower than its true value because children would appear to carry the same genotype for longer. Re-colonizations, if they happen, would be more frequent when the number of carriers is higher. Pneumococcal carriage is much more frequent in Portuguese DCC [Mato et al.2005] than in Finnish DCC [Syrjänen et al.2001], which could account for an apparent clearance rate to be estimated lower in Portuguese day-care children than in Finnish day-care children. Taking re-colonizations into account might result in correlation problems similar to those observed here with the community acquisition rate and the within-group transmission parameter. The model might not be able to distinguish between high rates of re-colonization and low rates of clearance or the opposite.

In this study, strains were identified using the serotype, genotype and antiobiotic resistance pattern that were used to define clones. This might be more accurate than using only the serotype information, as in the Finnish study [Hoti et al.2009]. The exact mechanism of competition would determine the effect of using different levels of strain identification. If strains compete based on the capsular type, then using more information than the serotype would lead to distinguishing strains that do not compete with each other and result in an under-estimation the competition parameter ϕ. If the competition between strains depends on something other than the serotype and only the serotype information is used, then the strains will not be adequatly distinguished and some super-colonizations would not be noticed which would also affect the parameter estimations. Unfortunately the competition mechanism behind pneumococcal competition in the nasopharynx is poorly understood [Auranen et al.2010].

The assumption that children did not undergo more than one transition in one month did not hold. Simulated data made it possible to see that sometimes children can undergo two transitions in very short periods of time. Using a daily sample over 10 months, it was possible to obtain accurate estimates, centered around the value that had been chosen for the simulation. However, this is not a realistic sampling strategy. Aside from the heavy work load and high cost, such frequent samplings could affect the nasopharyngeal flora and bias the results. This model is therefore impractical because of the assumption that one can count the number of transitions from the state of the children at two consecutive sampling times. This estimation model depends

on that number to determine the posterior probabilities for the parameters.

An alternative to increasing sampling frequency would be to define a likelihood model based on event history analysis, using transition times as observed data. These transition times would need to be inferred, for example using Bayesian latent process approach, where another MCMC (Markov Chain Monte Carlo) algorythm would sample the space of possible carriage histories consistent with the model and the states at the sampling times.

[Andersen et al.1997] Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1997). *Statistical Models Based on Counting Processes*. Springer, corrected edition.

[Auranen et al.2010] Auranen, K., Mehtälä, J., Tanskanen, A., and Kaltoft, M. S. (2010). Between-Strain competition in acquisition and clearance of pneumococcal Carriage—Epidemiologic evidence from a longitudinal study of Day-Care children. *American Journal of Epidemiology*, 171(2):169 –176.

[Bogaert et al.2004] Bogaert, D., de Groot, R., and Hermans, P. (2004). Streptococcus pneumoniae colonisation: the key to pneumococcal disease. *The Lancet Infectious Diseases*, 4(3):144–154.

[Brueggemann et al.2003] Brueggemann, A. B., Griffiths, D. T., Meats, E., Peto, T., Crook, D. W., and Spratt, B. G. (2003). Clonal relationships between invasive and carriage streptococcus pneumoniae and serotype- and Clone-Specific differences in invasive disease potential. *The Journal of Infectious Diseases*, 187(9):1424–1432. ArticleType: research-article / Full publication date: May 1, 2003 / Copyright © 2003 The University of Chicago Press.

[Centers for Disease Control and Prevention2000] Centers for Disease Control and Prevention (2000). Preventing pneumococcal disease among infants and young children. recommendations of the advisory committee on immunization practices (ACIP). *Recommendations and Reports: Morbidity and Mortality Weekly Report.*, 49(RR-9):1–35. PMID: 11055835.

[Coelho2009] Coelho, F. (2009). bayesian-inference - project hosting on google code. http://code.google.com/p/bayesian-inference/.

[Gelman et al.1995] Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman & Hall.

[Gelman and Rubin1992] Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472. ArticleType: research-article / Full publication date: Nov., 1992 / Copyright © 1992 Institute of Mathematical Statistics.

[Gillespie1977] Gillespie, D. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2361, 2340.

[Hoti et al.2009] Hoti, F., Erasto, P., Leino, T., and Auranen, K. (2009). Outbreaks of streptococcus pneumoniae carriage in day care cohorts in finland - implications for elimination of transmission. *BMC Infectious Diseases*, 9(1):102.

[Keeling and Rohani2008] Keeling, M. J. and Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton University Press.

[Klugman2008] Klugman, K. P. (2008). *Pneumococcal vaccines: the impact of conjugate vaccine*. ASM Press.

[Lipsitch1997] Lipsitch, M. (1997). Vaccination against colonizing bacteria with multiple serotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 94(12):6571 –6576.

[Lipsitch et al.2000] Lipsitch, M., Dykes, J. K., Johnson, S. E., Ades, E. W., King, J., Briles, D. E., and Carlone, G. M. (2000). Competition among streptococcus pneumoniae for intranasal colonization in a mouse model. *Vaccine*, 18(25):2895–2901.

[Mato et al.2005] Mato, R., Sanches, I. S., Simas, C., Nunes, S., Carriço, J. A., Sousa, N. G., ao, N. F., Saldanha, J., Brito-Avô, A., Almeida, J. S., and Lencastre, H. D. (2005). Natural history of drug-resistant clones of streptococcus pneumoniae colonizing healthy children in portugal. *Microbial Drug Resistance (Larchmont, N.Y.)*, 11(4):309–322. PMID: 16359190.

[May and Nowak1994] May, R. M. and Nowak, M. A. (1994). Superinfection, metapopulation dynamics, and the evolution of diversity. *Journal of Theoretical Biology*, 170(1):95–114. PMID: 7967636.

[O'Brien and Nohynek2003] O'Brien, K. L. and Nohynek, H. (2003). Report from a WHO working group: standard method for detecting upper respiratory carriage of streptococcus

pneumoniae. *The Pediatric Infectious Disease Journal*, 22(2):e1–11. PMID: 12586987.

[Patil et al.2010] Patil, A., Huard, D., and Fonnesbeck, C. (2010). PyMC: Bayesian stochastic modelling in python. *Journal of Statistical Software*, 35(4):1–81.

[Revai et al.2006] Revai, K., McCormick, D. P., Patel, J., Grady, J. J., Saeed, K., and Chonmaitree, T. (2006). Effect of pneumococcal conjugate vaccine on nasopharyngeal bacterial colonization during acute otitis media. *Pediatrics*, 117(5):1823–1829. Vaccine, Otitis, Colonization.

[Ruoff et al.2003] Ruoff, K., Whiley, R. A., Beighton, D., Murray, P. R., Baron, E. J., Jorgensen, J. H., Pfaller, M. A., and Yolken, R. H. (2003). Streptococcus. In *Manual of clinical microbiology*, pages 405–421. American Society for Microbiology, 8th edition.

[Sá-Leão et al.2008] Sá-Leão, R., Nunes, S., Brito-Avô, A., Alves, C. R., ao A. Carriço, J., Saldanha, J., Almeida, J. S., Santos-Sanches, I., and de Lencastre, H. (2008). High rates of transmission of and colonization by streptococcus pneumoniae and haemophilus influenzae within a day care center revealed in a longitudinal study. *Journal of Clinical Microbiology*, 46(1):225–234. PMID: 18003797 PMCID: 2224302.

[Salvatier2010] Salvatier, J. (2010). Python package index : multichain_mcmc 0.3. http://pypi.python.org/pypi/multichain_mcmc/0.3.

[Syrjänen et al.2001] Syrjänen, R. K., Kilpi, T. M., Kaijalainen, T. H., Herva, E. E., and Takala, A. K. (2001). Nasopharyngeal carriage of streptococcus pneumoniae in finnish children younger than 2 years old. *The Journal of Infectious Diseases*, 184(4):451–459. ArticleType: research-article / Full publication date: Aug. 15, 2001 / Copyright © 2001 The University of Chicago Press.

[Vrugt et al.2009] Vrugt, J. A., Braak, C. J., Diks, C. G., Robinson, B. A., Hyman, J. M., and Higdon, D. (2009). Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. 10(3):273–290.

[Weinberger et al.2009] Weinberger, D. M., Trzciński, K., Lu, Y., Bogaert, D., Brandes, A., Galagan, J., Anderson, P. W., Malley, R., and Lipsitch, M. (2009). Pneumococcal capsular polysaccharide structure predicts serotype prevalence. *PLoS Pathog*, 5(6):e1000476.

[Zhang et al.2004] Zhang, Y., Auranen, K., and Eichner, M. (2004). The influence of competition and vaccination on the coexistence of two pneumococcal serotypes. *Epidemiology and Infection*, 132(6):1073–1081. PMID: 15635964.