Discrete-time observations of *Streptococcus pneumoniae* colonisation – analysis and design under continuous-time Markov models

Juha Mehtälä

Department of Vaccination and Immune Protection
National Institute for Health and Welfare
and
Department of Mathematics and Statistics
University of Helsinki

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public examination in Auditorium XV, University main building, on 17$^{\text{th}}$ of June 2015 at 12 o'clock noon.

Helsinki 2015

ISBN 978-951-51-1253-8 (nid.)
ISBN 978-951-51-1254-5 (PDF)
http://ethesis.helsinki.fi/

Printed at Unigrafia

Helsinki 2015

Supervised by
Docent Kari Auranen, National Institute for Health and Welfare and University of Helsinki

Docent Sangita Kulathinal, University of Helsinki

Reviewed by
Professor Tom Britton, Stockholm University

Professor Esa Läärä, University of Oulu

Opponent
Professor Elizabeth Halloran, Fred Hutchinson Cancer Research Center and University of Washington

Custos
Professor Jukka Corander, University of Helsinki

Abstract

Continuous-time Markov processes with a finite state space can be used to model countless real world phenomena. Therefore, researchers often encounter the problem of estimating the transition rates that govern the dynamics of such processes. Ideally, the estimation of transition rates would be based on observed transition times between the states in the model, i.e., on continuous-time observation of the process. However, in many practical applications only the current status of the process can be observed on a pre-defined set of time points (discrete-time observations).

The estimation of transition rates is considerably more challenging when based on discrete-time data as compared to continuous observation. The difficulty arises from missing data due to the unknown evolution of the process between the actual observation times. To be able to estimate the rates reliably, additional constraints on how they vary in time will usually be necessary.

A real world application considered in this thesis involves the asymptomatic carriage state (colonisation) with the bacterium *Streptococcus pneumoniae* (the pneumococcus). The pneumococcus has over 90 strains and for understanding the dynamics of the pneumococcus among humans it is important to understand within-host competition between these strains. Research questions regarding competition in this thesis are: does colonisation by one serotype protect from acquisition of other serotypes and is clearance affected by concurrent colonisation by other serotypes? A question regarding the implication of competition to pneumococcal dynamics after vaccination is also of interest. In addition, vaccine protection may be heterogeneous across individuals, leading to a question about how well such vaccine protection can be estimated from discrete-time data.

When only discrete-time observations are available, the decision when to measure the current status of the process is particularly important. With measurements that are temporally apart from each other, information about the state of the process at one point does not give information about the state at the other points. When measurements are very close to each other, knowing the state at one point bears information about the state at other, temporally close points.

This thesis addresses the estimation of transition rates based on repeated observations of the current status of an underlying continuous-time Markov process. Applications to actual data concern the process of pneumococcal colonisation. Optimal study designs are considered for improved future studies of similar type, applications including but not limited to pneumococcal colonisation studies.

List of original publications

This dissertation rests on the five original articles listed below and referred to in the text by their Roman numerals. Article summaries can be found at the end of this thesis.

I Between-strain competition in acquisition and clearance of pneumococcal carriage – epidemiologic evidence from a longitudinal study of day-care children. Auranen K, Mehtälä J, Tanskanen A, Kaltoft MS. *American Journal of Epidemiology.* 2010;**171**(2):169-76.

II Competition between Streptococcus pneumoniae strains: implications for vaccine-induced replacement in colonization and disease. Mehtälä J, Antonio M, Kaltoft MS, O'Brien K, Auranen K. *Epidemiology.* 2013;**24**(4):522-9.

III Optimal designs for epidemiologic longitudinal studies with binary outcomes. Mehtälä J, Auranen K, Kulathinal S. *Statistical Methods in Medical Research.* 2011;DOI: 10.1177/0962280211430663.

IV Optimal observation times for multi-state Markov models – applications to pneumococcal colonisation studies. Mehtälä J, Auranen K, Kulathinal S. *Journal of the Royal Statistical Society: Series C.* 2015;**64**(3):451-68.

V Estimation and interpretation of heterogeneous vaccine efficacy against recurrent infections. Mehtälä J, Dagan R, Auranen K. Submitted for publication. 2015.

Author's contribution

I I implemented the statistical methods, performed the data-analyses and contributed to interpretation of the results.

II I implemented the statistical methods and performed the data-analyses. I wrote the paper with support from the co-authors.

III I implemented the methods and wrote the paper together with the co-authors.

IV I implemented the methods and wrote the paper with support from the co-authors.

V I implemented the methods and was primarily responsible for writing the paper.

Acknowledgements

My work involving the application of statistical and mathematical methods to infectious diseases started in June 2007, with an internship at the National Public Health Institute, later to be known as the National Institute for Health and Welfare. Back then, I joined a statistical modelling group that was part of the PneumoCarr project, a five year long work that had started in 2006. The modelling group consisted of researchers Kari Auranen, Panu Erästö, Fabian Hoti, Hanna Rinta-Kokko and Markku Nurhonen, all of whom I am grateful for useful discussions. After the internship I continued the work as a part time assistant. Eventually, after suggestions from my supervisors Kari Auranen and Sangita Kulathinal, that work led to writing this thesis. For Kari I would like to relay special thanks in guiding me in scientific writing. For my other supervisor, Sangita Kulathinal, I am especially thankful about introducing me to the world of optimisation of study designs.

I am very grateful to the two pre-examiners, Professors Tom Britton and Esa Läärä, who took the time and effort to review my thesis and suggested how to further improve it. I am most grateful to Professor Elizabeth Halloran who accepted the invitation to be my opponent.

The PneumoCarr project, of which part of my work was funded as well, was supported by the grant (37875) from the Bill and Melinda Gates foundation. I would also like to thank the Finnish Academy of Sciences for a grant (number 138068) that allowed continuing the initiated work.

Contents

1 Introduction

The application of statistics in medicine has its foundations in probability theory, greatly influenced by Pierre de Fermat (1601 - 1665) and Blaise Pascal (1623 - 1662), and in the general principles of statistical inference with significant contributions by Carl Friedrich Gauss (1777 - 1855), Thomas Bayes (1702 - 1761), Ronald Fisher (1890 - 1962), and Jerzy Neyman (1894 - 1981). One of the earliest who applied statistical methods in medicine was John Snow (1818 - 1858) whose spatial analysis of cholera cases revealed that the central source of this disease was a water well [1]. Since the times of John Snow, the use of statistics in medicine and epidemiology has dramatically increased up to a point in which one of the most influential medical journals chose statistics in medicine as the 4^{th} most important medical milestone in the past millennium [2].

Dynamic modelling is increasingly popular in epidemiology, particularly in describing the spread of infectious diseases [3, 4]. The history of using these models dates as far back as the early 20^{th} century when Ronald Ross described malaria transmission by using difference equations [5, 6]. The models introduced by Ross were further investigated and solved by Alfred Lotka [7, 8]. One of the major conclusions in these works was the notion of a critical mosquito density below which malaria transmission would not occur. A little later, based largely on Ross' pioneering investigations, Kermack and McKendrick used differential equations to understand observed patterns of plague epidemics [9]. They used an SIR model that describes individuals as either susceptible (S), infected (I) or recovered (R) in a closed homogeneously mixing population and the particular interest was to determine a threshold (as a function of population size) below which the epidemic dies out. George MacDonald continued the work of Ross and Lotka, emphasising the same threshold quantity, the basic reproduction number, nowadays commonly denoted as R_0 [10, 11]. The analyses of R_0 also led to practical actions in the form of exterminating mosquitoes with dichlorodiphenyltrichloroethane (DDT) in order to eliminate malaria transmission. One of the most common modern examples of interventions aimed to control infectious diseases are vaccinations. The evaluation of vaccine efficacy and its relation to the threshold under which the infection will die out is made almost as a matter of routine. Dynamic models have an important role in these considerations as they help to understand the long term behaviour of infectious diseases in vaccinated populations.

There is also a long history of using statistical methods in analysing infectious disease data [12, 13, 14]. An important example are vaccine studies in which the aim is to estimate the protective vaccine efficacy, usually as the relative reduction in the risk of infection among the vaccinated in comparison to the unvaccinated [15]. Another example is model-based estimation that does not rely directly on statistical associations but gives information about rates of acquisition, transmission or durations of latent and infectious period of an infection. However, when employing such estimates, it is particularly important that the model parameters have correct interpretations, their estimators are unbiased and have low uncertainties. Consequently, the usability of inference results depends substantially on the model specification, the methods of statistical inference and on the study design, i.e., how the data were acquired. For instance, when modelling the spread of an infectious disease in a population using differential equations, one usually needs a set of constants that are taken as input parameters. It is common to adopt these constants as model-based estimates resulting from statistical analysis[16, 17]. Hence, any such input parameters would need to have a similar interpretation as in the source study, which may be problematic because of unadjusted confounding or different conditioning on influential

background variables. Parameter estimators may also have too high uncertainty to make strong conclusions. With better design of studies, uncertainty in parameter estimates can be reduced allowing firmer conclusions to be made.

Early works which emphasised the importance of study designs are The Arrangement of Field Experiments (1926) and The Design of Experiments (1935) by Fisher. These introduced ideas such as block designs of which one example is the allocation of agricultural treatments randomly into blocks such that each treatment occurs exactly once in each row and in each column (Latin square). Using this design instead of completely random allocation leads to smaller variances for the treatment estimators. Indeed, the minimisation of the variances of the parameter estimators of interest is a common desideratum in study design considerations. When the model parameters form a vector, the covariance matrix is often evaluated as the inverse of the Fisher information, and the minimisation of the elements of this matrix can be performed in many ways. Based on the choice of how the multi-dimensional variance is minimised, different optimality criteria have been defined. Examples are the E-optimality that seeks to minimise the variance of the worst estimated linear combination of the parameters by maximising the minimum eigenvalue of the Fisher information matrix, and D-optimality that seeks to minimise the volume of the variance ellipsoid by maximising the determinant. Major contributors to these definitions and on early optimisation theory in general include Jack Kiefer (1924-1981), Abraham Wald (1902-1950) and Jacob Wolfowitz (1910-1981).

Another significant contributor to the research of optimal designs, particularly under linear regression models, was Gustav Elfving (1908-1984). He considered the optimal allocation of the independent variable X_i in estimating parameters α or β in the model $Y_i = \alpha + X_i\beta + \eta$, based on observed Y_i. The solution proposed for both estimation problems is that one should only choose the minimum and maximum values of X_i, but the proportion in which these observations are made depends on whether the aim is to estimate α or β. Elfving's solution to the linear design problem does not depend on the values of the model parameters α and β. In many design problems, however, this is not the case. In classical approaches it is then common to use a guess for the model parameters to derive a locally optimal design. In the Bayesian setting, it is natural to integrate over a parameter range with respect to a prior distribution.

This thesis is about study designs and statistical inference motivated by applications to infectious diseases. In particular, all applications considered in this thesis involve a bacterium *Streptococcus pneumoniae* (the pneumococcus) in two alternative roles. On one hand, the pneumococcus has motivated important research questions, which ultimately can be related to disease burden in humans, and statistical methods were employed to answer these questions. On the other hand, there are several questions related to the pneumococcus that still require further investigations. In this thesis, design guidelines are derived to improve longitudinal pneumococcal studies in the future. Specifically, studies considered here pertain to pneumococcal colonisation of the human nasopharynx, an asymptomatic state that precedes disease. Furthermore, in many other applications discrete-time data are collected in similar manner to learn the dynamics of continuous-time processes, so the presented design principles apply to other than pneumococcal colonisation studies as well.

1.1 Brief description of the pneumococcus

This section introduces some background information of the example pathogen of this thesis, the pneumococcus. The pneumococcus is a bacterium that causes high morbidity and mortality worldwide, especially among young children [18]. A pre-requisite for pneumococcal disease is colonisation of the upper back throat (nasopharynx) [19]. Colonisation of the nasopharynx is mostly asymptomatic and common in relation to disease that results when the pneumococcous invades other tissues such as the lungs or bloodstream leading to a potentially life-threatening condition.

The ultimate goal regarding the pneumococcus is to reduce the burden of disease it causes worldwide. This has proven to be challenging, much stemming from the multi-strain nature of the bacterium. The pneumococcus has over 90 subtypes (serotypes or strains) that differ by their polysaccharide capsule. The capsule lies outside the cell wall and defines many properties related to the bacterium's interaction with the environment, e.g. with the human immune system. A particular example of an attempt to reduce pneumococcal disease burden is vaccination. Indeed, pneucmoccoccal conjugate vaccines have proven to be efficacious in preventing the most serious forms of disease caused by serotypes included in the vaccines. However, the current vaccines include antigenic components against only $7 - 13$ serotypes. As pneumococcal conjugate vaccination also reduces nasopharyngeal colonisation with the vaccine serotypes, it opens a niche for the non-vaccine types leading to potential replacement in disease as well [20, 21]. The phenomenon of replacement reflects the fact that pneumococcal serotypes do not colonise the human host independently but actually compete against each other to occupy their own niche.

Because colonisation is common in relation to disease and the main source of transmission between humans, it provides good premises to study some important characteristics of the pneumococcus. These include the dynamics of colonisation in a population, risk factors for acquiring colonisation (and thus perhaps disease) and within-host competition between different serotypes. As pneumococcal colonisation is considered a necessary step in the pathogenesis of pneumococcal disease, there may be a rationale to use colonisation in the evaluation of new vaccines as well [22].

Colonisation is usually detected by placing a calcium alginate swab to the nasopharynx, cultivating the sample on an agar plate, and if found positive for colonisation, the serotype or serotypes in question can be subsequently determined [23]. The standard method to detect pneumococcal colonisation has a high sensitivity to indicate whether a host is colonised. However, in the case of many serotypes simultaneously colonising the same host, conventional methods do not necessarily perform well in detecting more than one of these types.

1.2 Pneumococcal colonisation data and general aspects of inference

The process of pneumococcus colonising the human host is often modelled using a Markov model with a finite state space [24, 25, 26]. This class of statistical models is common to all work in this thesis while similar models have a number of other applications as well. Although bacterial colonisation may not strictly be an on/off process, the state of an individual can be thought to change from non-colonised to colonised after the bacterial growth reaches a certain biologically significant threshold. Hence, modelling with a finite state space is likely to be sufficient for many purposes.

Regarding applications to pneumococcal colonisation, SIS (susceptible-infected-susceptible) models in which the infection can be acquired multiple times have been used (e.g. [25, 26]). This has been based on the fact that innate or acquired immunity do not completely prevent new acquisitions. The simplest model for an infection that can be acquired multiple times is the basic SIS model with two states that describe whether the individual is infected or susceptible. In our example case, two states do not fully describe the appropriate dynamics due to the multi-strain nature of the pneumococcus. In addition, it is possible that two or more strains colonise the same host simultaneously. In practice, common serotypes are often modelled as individual strains while the rest are pooled into a combined state of rarer serotypes. Simultaneous colonisation with two different types may be included as its own state if the measurements are sensitive enough to detect multiple colonisation. Simultaneous colonisation with more than two serotypes is rarely observed or included in the models of pneumococcal colonisation.

The dynamics of colonisation can be described in terms of rates of transitions between the model states. Typical research questions related to pneumococcal colonisation involve the estimation of transition rates, covariate effects on those rates, and ratios of rates between two groups with different treatment status. For instance, the baseline hazards of acquisition and clearance, the effect of exposure on acquisition, or comparison of acquisition rates between vaccinated and unvaccinated groups could be of interest. A particular motivation for this thesis arises from competition between pneumococcal serotypes. Here, competition is described as the relative rate of acquiring or clearing colonisation of a serotype when the host is simultaneously colonised with some other serotype as compared to the corresponding rates without the presence of any other serotype. Previous studies with observational data and this type of definition of competition are limited.

Under models in which the infection can be acquired multiple times it is not possible to base the estimation of transition dynamics on cross-sectional data. For instance, consider a two-state model that describes whether the infection is present or not. Cross-sectional data of the stationary prevalence of the infection are sufficient to estimate the ratio of acquisition and clearance rates of the infection. Alternatively, if one of the rates is fixed the other can be estimated [27]. The situation is similar when estimating rates in a comparative study based on two groups of subjects (e.g. vaccinated and unvaccinated). It is possible to estimate such effects from prevalence data by assuming that the effect is either solely on acquisition ($S \rightarrow I$) or clearance ($I \rightarrow S$).

In this thesis, statistical inference about the transition rates is based on repeated measurements of the current status of pneumococcal colonisation in the study subjects. Continuous-time observations are not available mainly because colonisation does not cause immediately detectable symptoms. Such asymptomatic nature is not exceptional, however. For instance, *Neisseria meningitidis* has a similar asymptomatic colonisation state [28]. A considerable proportion of influenza infections do not cause detectable symptoms either [29].

The current status data are interpretable as a collection of snapshots from the underlying continuous-time status of each individual. Between any two consecutive observations there is an unobserved part and the path that the process takes between the two observed states remains unobserved (i.e. missing). In particular, one can question whether the process visited other states during the unobserved period or made a direct transition from one observed state to another, or even remained in the same state the whole time. In addition, if a state change occurred, the times of these events are unknown. With a relatively short time in between the observations, it is likely that only few state changes

remain unobserved while samples lying temporally apart from each other leave more room for unobserved events.

When the likelihood of the model parameters under a continuous-time process is based on current status data, the possibility of unobserved events between the observations needs to be taken into account. A common approach in applied epidemiology is to assume a minimum number of transitions (often only one) that makes the observations possible to occur somewhere between (e.g. midway) the two observed states [30]. This approach is applied also in Article I. Another extreme is complete treatment of the unobserved part by integration over all possible events [31]. The difficulty of calculating such an integral depends on how complex time-dependence is allowed for the transition rates. In the simplest case the transition rates are assumed to be constant. Transition probabilities that take into account all possible transition paths can then be derived easily [32]. Otherwise, numerical integration methods have been applied [31]. In Articles II − V it is assumed that the transition rates are constant in time.

The investigator often has many alternative options to allocate repeated observations in a study. The optimal allocation of observation times then becomes an important design question so as to improve how well the study answers to the posed research questions. For instance, if the problem is to estimate the parameters of a statistical model, one can ask how the time spacing between consecutive samples affects the bias and variance of the parameter estimators. With a short time spacing, repeated samples are typically more correlated than with a long one. With long enough time spacing, observations may arise almost independently, given the underlying process. It has not been comprehensively studied how the time spacing and thus the correlation between consecutive samples should be to obtain unbiased estimators with as small a variance as possible for the quantities of interest. The main design question considered in this thesis is to determine which time spacing between zero (highly dependent samples) and infinity (independent samples) is optimal in to estimate parameters of interest. Design question related to time spacings between consecutive observations are considered in Articles III, IV and V. Other study design questions including the total number of samples to be collected, the number of individuals versus repeated observations from the same individuals and the initial state of the process to collect the first samples are considered in Article III.

Measurements of the current infection status may contain errors or imprecisions. For instance, because of sampling methods, the infected status may sometimes be incorrectly observed as susceptible. The type of infection may also be incompletely identified. With pneumococcal colonisation, a particular problem arises when more than one subtype colonise the same host simultaneously. It is challenging to determine all subtypes present in the nasopharynx. The analysis of within-host competition between different serotypes is then particularly problematic. In Article II, a hidden Markov model for such incomplete observations is considered, and it is studied how the analyses about competition are affected by the sensitivity to detect simultaneously colonising subtypes.

2 Markov models

Let $\{\zeta_i(\tau); \tau \geq 0\}$, be a continuous-time stochastic process with a finite state space $S = \{1, ..., n\}$. The process $\zeta_i(\tau)$ describes the state of an individual i at time τ. For instance, with $S = \{1, 2\}$, $\zeta_i(\tau)$ could denote whether the individual is susceptible (1) or infected (2) at time τ. In epidemiology, models of this type are often referred to as compartmental.

The dynamics of the process $\zeta_i(\tau)$ can be defined in terms of its transition rates:

$$\lambda_{j,k}(\tau) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(\zeta_i(\tau + \Delta t) = k \mid \zeta_i(\tau) = j)}{\Delta t},$$

where $j, k \in S$, $j \neq k$, and $\lambda_{k,k} = -\sum_{j=1, j\neq k}^{n} \lambda_{k,j}$. The rates can be represented as an $n \times n$ matrix $\mathbf{q}(\tau)$, where the $(j,k)^{\text{th}}$ element is $\lambda_{j,k}(\tau)$. In this thesis, the transition rates are not allowed to depend on the history of the process but only on the current status at time τ, i.e. only Markov processes are considered. Further introduction to Markov processes and their applications can be found e.g. from [33] and [34].

The transition probability that, conditioned on being in state j at time τ, individual i is in state k after time t has elapsed is denoted by

$$\mathbf{p}^t_{j,k}(\tau) = \mathbb{P}(\zeta_i(\tau + t) = k \mid \zeta_i(\tau) = j).$$

The transition probability matrix with the $(j,k)^{\text{th}}$ element $\mathbf{p}^t_{j,k}(\tau)$ is denoted by $\mathbf{p}^t(\tau)$. The transition rate matrix generates the transition probabilities for all time intervals t. In particular, the relation between the transition rates and the transition probabilities satisfies the Kolmogorov forward equations:

$$\frac{d\mathbf{p}^t(\tau)}{dt} = \mathbf{p}^t(\tau)\mathbf{q}(\tau),$$

with the initial condition $\mathbf{p}^0(\tau) = \boldsymbol{I}_n$, the $n \times n$ identity matrix. The solution of these equations can be written as a product integral as follows. Given a partitioning of the interval $(\tau, \tau + t)$ into m subintervals according to points $\tau = t^1 < t^2 < ... < t^m = \tau + t$, the transition probability matrix is

$$\mathbf{p}^t(\tau) = \lim_{\max |t^i - t^{i-1}| \to 0} \prod_{i=1}^{m} \Big(\boldsymbol{I} + d\mathbf{p}(t^i) \Big),$$

where m increases as $\max |t^i - t^{i-1}| \to 0$. The term $\boldsymbol{I} + d\mathbf{p}(t^i)$ is the transition probability matrix over a short time interval (t^{i-1}, t^i) and $d\mathbf{p}(t^i)$ approaches $|t^i - t^{i-1}| \times \mathbf{q}(t^i)$ as $|t^i - t^{i-1}| \to 0$. The above considerations pose no assumptions about how the transition rates vary in time. However, without further assumptions, the product integral formulation does not easily lend to calculating the transition probabilities unless continuous observations of the process are available.

2.1 Inference with continuous-time data

If the process is observed continuously in time, one can estimate the cumulative transition rates (i.e. the elements $d\mathbf{p}(t^i)$) non-parametrically with the Nelson-Aalen estimator [35, 36]. The empirical transition probability matrix for $\mathbf{p}^t(\tau)$ is then the Aalen-Johansen estimator, obtained as the product integral of the Nelson-Aalen estimator [37, 38, 39]. The Nelson-Aalen and Aalen-Johansen estimators make no restricting assumptions about how the rates vary over time.

Parametric approaches based on the likelihood for continuous-time observations are lucidly presented by Andersen and Keiding [40] and more rigorously in [14]. In brief, suppose that N individuals are observed for time intervals $[0, \tau_i], i = 1, ..., N$. The probability density that, conditional on being in state h at time s, one stays in that state until time $t > s$

and then moves to another state j at time t is $p_{h,j}(s,t) = \lambda_{h,j}(t)\exp(-\int_s^t \sum_{k\neq h}\lambda_{h,k}(u)du)$. Let $N^i_{h,j}(\tau_i)$ denote the number of observed transitions for individual i from state h to j by during time interval $[0,\tau_i]$. The corresponding event times are denoted as $T^{1,i}_{h,j} < ... < T^{N^i_{h,j}(\tau_i),i}_{h,j}$. The likelihood function of rates $\boldsymbol{\lambda}$ based on continuous-time data $\boldsymbol{Y}$ from N independent individuals is

$$L(\boldsymbol{\lambda};\boldsymbol{Y}) = \prod_{i=1}^{N}\pi_0^i \prod_{h=1}^{n}\prod_{j\neq h}\prod_{k=1}^{N^i_{h,j}(\tau_i)} \lambda_{h,j}(T^{k,i}_{h,j})\exp\Big(-\int_0^{\tau_i}\lambda_{h,j}(u)\mathbf{1}^i_h(u)du\Big), \tag{1}$$

where for individual i, π_0^i is the probability of being in the observed state at time 0, and $\mathbf{1}^i_h(u)$ the indicator of being in state h at time $u-$, just before time u.

2.2 Likelihood based on current status data

This thesis considers situations in which only current status data are available. With repeated measurements of each individual's status, this is equivalent to observing the frequency of transitions between the model states over fixed time intervals. The interest is to estimate the rate matrix $\mathbf{q}(\tau)$ based on these frequencies. For this kind of inference to be possible, transition probabilities need to be expressed in terms of the transition rates, which usually requires restricting assumptions about how the rates vary in time.

For a formal presentation, denote the number of individuals by N, the number of samples collected from individual i by K_i and the observation times by $\tau_i^k,\ k = 1,...,K_i; i = 1,...,N$. Denote the observed data by $\boldsymbol{Y} = \{Y_i(\tau_i^k); k = 1,...K_i; i = 1,...,N\}$, where $Y_i(\tau_i^k)$ is the state of individual i at time τ_i^k. If individuals are statistically independent, the joint probability of these data under the Markovian assumption is

$$\prod_{i=1}^{N}\mathbb{P}(\zeta_i(\tau_i^1) = Y_i(\tau_i^1))\prod_{k=2}^{K_i}\mathbb{P}(\zeta_i(\tau_i^k) = Y_i(\tau_i^k)|\zeta_i(\tau_i^{k-1}) = Y_i(\tau_i^{k-1})), \tag{2}$$

where the initial distribution $\mathbb{P}(\zeta_i(\tau_i^1) = Y_i(\tau_i^1))$ usually needs its own specification. The probability $\mathbb{P}(\zeta_i(\tau_i^k) = Y_i(\tau_i^k)|\zeta_i(\tau_i^{k-1}) = Y_i(\tau_i^{k-1}))$ is for the transition that the individual i is in state $Y_i(\tau_i^k)$ at time τ_i^k, conditional on being in state $Y_i(\tau_i^{k-1})$ at time τ_i^{k-1}. Here, as always in this thesis, it is assumed that there exists a continuous-time process that induces the transition probabilities, that is, the process is defined by a rate matrix $\mathbf{q}(\tau)$. To emphasise this connection, the transition probability will be denoted as the corresponding element of the transition probability matrix $\mathbf{p}(Y_i(\tau_i^{k-1}), Y_i(\tau_i^k))$. The transition probability matrix is called embeddable if there exists a generating rate matrix. Although it is assumed that such a rate matrix exists in theory, the sample-based transition frequency matrix may not always be embeddable [41, 42].

In the following, the rate matrix $\mathbf{q}$ is defined through a parameter vector $\boldsymbol{\Theta}$, but the dependence of $\mathbf{q}$ and $\mathbf{p}$ on $\boldsymbol{\Theta}$ is omitted from the notation for brevity. It also assumed that the transition rates are constant or piecewise constant in time, so that the induced transition probability matrix can be calculated from the rates. In particular, when $\mathbf{q}(\tau) = \mathbf{q}$ for all τ, the transition probability matrix for a time interval of length t is $\mathbf{p}^t = \exp(\mathbf{q}\times t)$, where exp refers to the matrix exponential function ($\exp(\mathbf{q}\times t) = \sum_{i=0}^{\infty}((\mathbf{q}\times t)^i)/i!$) [43]. The probability of data described by equation (2) with the initial distribution π provides

the likelihood function of the parameters $\boldsymbol{\Theta}$:

$$L(\boldsymbol{\Theta}; \boldsymbol{Y}) = \prod_{i=1}^{N} \pi(Y_i(\tau_i^1)) \prod_{k=2}^{K_i} \mathbf{p}(Y_i(\tau_i^{k-1}), Y_i(\tau_i^k)).$$

A common assumption in modelling the spread of an infection is that the transition rates for a particular individual depend not only on his or her current status but also on the status of other individuals. In particular, the rate of acquiring the infection is often assumed to depend on the current number of infected individuals in the surrounding population. The state space of the process then needs to entail the joint status of all individuals. For a population of N individuals, let the process $\boldsymbol{\zeta}(\tau)$ carry the information of each individual's status at time τ. The dimension of the joint state space of the process $\boldsymbol{\zeta}(\tau)$ is n^N. Because the current status of the population carries information about the current number of infected individuals, the process is still Markovian. In particular, it is governed by a transition rate matrix $\mathbf{Q}$ of dimension $n^N \times n^N$. The transition probability matrix for the process $\boldsymbol{\zeta}(\tau)$ is $\mathbf{P}^t = \exp(\mathbf{Q})$.

In order to write the likelihood for parameters $\boldsymbol{\Theta}$ that define $\mathbf{Q}$, let $\{\tau_i^k; j = 1, ..., K_i; i = 1, ..., N\} := \{\tau^j; j = 1, ..., l\}$ be the set of distinct observation times from all individuals in the study population. In case all individuals have not been observed at the same time points, the process $\boldsymbol{\zeta}(\tau)$ is only partially observed at least at some time points in the set $\{\tau^j\}$. Regarding the status of individual i at point $\tau^j, j = 1, ..., l$, denote the state by $Y_i(\tau^j)$ if it is observed and by $x_i(\tau^j)$ if it is missing. Let $R_i(\tau^j)$ be an indicator function for whether the status of individual i at τ^j is observed. For some realisation of the missing data $\boldsymbol{x}$, denote $\boldsymbol{Y}^{\boldsymbol{x}}(\tau^j) = Y_i(\tau^j)$ if $R_i(\tau^j) = 1$ and $\boldsymbol{Y}^{\boldsymbol{x}}(\tau^j) = x_i(\tau^j)$ if $R_i(\tau^j) = 0$. The likelihood based on the observed data $\boldsymbol{Y}$ is obtained by summing over all possible realisations of missing data $\boldsymbol{x}$:

$$L(\boldsymbol{\Theta}; \boldsymbol{Y}) = \sum_{\boldsymbol{x}} \pi(\boldsymbol{Y}^{\boldsymbol{x}}(\tau^1)) \prod_{k=2}^{l} \mathbf{P}(\boldsymbol{Y}^{\boldsymbol{x}}(\tau^{k-1}), \boldsymbol{Y}^{\boldsymbol{x}}(\tau^k)).$$

The choice to model discrete-time data using a continuous-time process is deliberate. It would have been possible to construct discrete-time transition probabilities instead of using a continuous-time process with its transition rates. However, an advantage of using continuous-time models is that they lead to estimates of transition rates rather than transition probabilities, which greatly facilitates the interpretation and comparison of results across different studies not necessarily based on the same sampling frequency. If studies were based on transition probabilities, rather than transition rates, transition probability matrices would need to be transformed to pertain to other time spacings. While transition probabilities for multiples of any given time spacing are obtained by matrix multiplication, other time spacings require taking matrix roots. The simplest condition for this to work is that the transition probability matrix ($\mathbf{p}$) is embeddable, i.e., the continuous-time process can be constructed based on observed transition probabilities [44]. In practice, this means that embeddability needs to be assumed in any case. Another argument favouring the use of transition rates is based on the need to model covariate effects. It is common to assume that the effects are multiplicative on the transition rates rather than on transition probabilities.

2.3 Models for missing data

When the process cannot be observed continuously in time, one has to confine to current status data observed at predefined time points. The status of the process in between any two observations can be described as missing by (study) design. The transition probability matrix for a continuous-time Markov process takes such missingness into account by definition. In the following, various other types of missingness are reviewed, together with a discussion about how they should be taken into account in making statistical inference.

Based on the mechanism by which the data are considered to remain unobserved, missing data can be categorised as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). These ideas originated from the seminal work of Rubin [45] and are still widely applicable [46]. To present formal definitions, recall that $R_i(\tau_i^j), j = 1, ..., K_i; i = 1, ..., N$, is the indicator for whether the status of individual i at time τ_i^j is observed. As before, let $\boldsymbol{y}$ denote the observed data and $\boldsymbol{x}$ the underlying but missing data. Missing completely at random is defined through condition $p(\boldsymbol{R}|\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{R})$, saying that the process that produces missing data is independent of the data, whether these are observed or not. The definition of missing at random (MAR) is also an independence condition. In particular, the condition for MAR is that missing data may depend on the observed part of the data but must not depend on the unobserved part: $p(\boldsymbol{R}|\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{R}|\boldsymbol{y})$. In the case of MNAR, whether or not data are observed depends on the data that remained unobserved.

Let $\boldsymbol{\xi}$ be a parameter vector that governs the process causing missing data. The correct likelihood for inferring about the model parameters, including ξ, is

$$L(\boldsymbol{\Theta}, \boldsymbol{\xi}; \boldsymbol{y}, \boldsymbol{R}) = p(\boldsymbol{y}, \boldsymbol{R}; \boldsymbol{\Theta}, \boldsymbol{\xi}) = \int_{\boldsymbol{x}} p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Theta}) p(\boldsymbol{R}|\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\xi}) d\boldsymbol{x}.$$

If the missing data are MCAR or MAR, likelihood-based inferences about $\boldsymbol{\Theta}$ can be based solely on $p(\boldsymbol{y}; \boldsymbol{\Theta}) = \int_{\boldsymbol{x}} p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Theta}) d\boldsymbol{x}$. This follows because under MCAR and MAR, $\boldsymbol{R}$ is independent of $\boldsymbol{x}$ and hence

$$\int_{\boldsymbol{x}} p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Theta}) p(\boldsymbol{R}|\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\xi}) d\boldsymbol{x} = p(\boldsymbol{R}|\boldsymbol{y}; \boldsymbol{\xi}) \int_{\boldsymbol{x}} p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Theta}) d\boldsymbol{x}.$$

In the case of MAR, other than strictly likelihood-based inferences may not be correct if based on $\int_{\boldsymbol{x}} p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Theta}) d\boldsymbol{x}$. For instance, estimators derived using generalised estimation equations require MCAR [47] whereas Bayesian methods require only MAR.

If the probability of not observing the status of an individual depends on the status itself (MNAR), a model for missing data need to be incorporated in the inferences. This is also the case if the measurement of state $Y_i(\tau_i^j)$ can result into an erroneous observation $Y_i'(\tau_i^j) \neq Y_i(\tau_i^j)$. Apart from Article II, data are assumed to be missing under MAR. In Article II, a model is constructed with an additional observational layer that takes into account the process of how the status of each individual is observed. The observational layer is needed because in the case of simultaneous colonisation of two pneumococcal serotypes it is possible that only one of these is observed. This type of missingness depends on the underlying state of the process and is thus MNAR. The model in Article II is a hidden Markov model and the likelihood is calculated recursively, similarly to the Viterbi algorithm [48].

A particular question addressed in Article II is the joint effect of the time spacing between consecutive measurements and the imperfect sensitivity to detect two simultaneously colonising serotypes (MNAR data) on the estimation of within-host competition

between pneumococcal strains. By simulation, it was shown for various time spacings how the analyses become biased if MNAR data are not properly accounted for. The simulation results were used to corroborate the unbiasedness of the findings about competition from the analysis of three data sets collected using varying sampling frequencies.

2.4 Parametrisation of transition rates

This section considers parametrisation of the rate matrix $\mathbf{q}(\tau)$ with a vector $\boldsymbol{\Theta}$ of small dimension. Research questions in this thesis are formulated as problems of estimating $\boldsymbol{\Theta}$. The dimension of the parameter space requires usually some restrictions as the following illustrates. In studies of pneumococcal colonisation, possible states include *susceptible* and *infected* by one of the strains (90 or more), totalling to more than 91 possible states. If all possible pairs of 90 strains are allowed to infect the same host simultaneously, there are $90 \times 89/2 = 4005$ additional states in the model. If all transition rates in a model with $n = 4096$ states were treated as distinct parameters, there would be $n(n-1)$ unknowns to be estimated.

One way of reducing the number of model parameters is to assume that certain transitions share the same rate. As a practical example, the average rate of clearance might be assumed to be the same among some subset of strains. The corresponding clearance rates could then be modelled using a single model parameter. In Articles I and II, relative rates, comparing already colonised and susceptible individuals, were parametrised in terms of a single parameter across all serotypes. For acquisition rates, mechanistic assumptions can be used to reduce the number of model parameters. For instance, a common transmission rate parameter (β) may be assumed across many strains, scaled for a specific strain according to exposure, i.e., the number of individuals infected with that strain (n_i). Acquisition rates in such cases are of the form $n_i\beta$. In a large population where the number of infected individuals is approximately constant in time (steady state) the whole term $(n_i/N)\beta$ can be modelled with a single parameter.

With regard to exposure, in Article II three different data sets on pneumococcal colonisation are employed assuming constant transition rates because information about exposure is not available in all three data sets. It is natural to rise a question of possible confounding since the research questions consider infectious diseases. Here the aim is to study between-strain competition defined through rate ratios such as the relative rate of acquisition of strain A when the host is currently colonised with another strain, as compared to acquisition (of A) when the host is non-colonised. Exposure to the strain would be a confounding factor if it is associated with the distribution of the time at-risk that is spend non-colonised and colonised with other strains that A itself. However, such associations were not found in the observed data.

Another way to reduce the number of model parameters is state aggregation, meaning that a group of states is considered as being one pooled state in the model [49]. Regarding applications to pneumococcal colonisation, commonly observed serotypes are often treated as their own states while the rest comprise an aggregated state of rarer serotypes (cf. Articles I and II). This is because the data are often insufficient for the estimation of strain-specific rates for the rarely observed types. Another possible motivation for aggregation is that the actual interest regards the estimation of transition rates between groups of states. For instance, the interest could be to estimate the total acquisition or clearance rates of all serotypes, or to estimate the acquisition and clearance rates for two pools of types, defined according to whether or not the vaccine can provide protection against

individual serotypes within the pools. Respectively, aggregated models with two and three states would then be appropriate. This type of aggregation is used in data analyses of Article V although the underlying model is constructed for a large number of states. By simulation, theoretical aggregate measures of transition rates are compared to those obtained by estimation based on aggregate level data.

In Article V, another question of parametrisation regards vaccine effects. In particular, Article V considers vaccines that may yield protection against a subset of all strains in the model and vaccine effects may also have differences at the individual level (random effects model). The strain-specific average vaccine effects were allowed to have their own parameters, but when conditioning on any one individual, the random effect was restricted to be equal for each strain. In addition, aggregate level vaccine-efficacy measures are considered to limit the number of estimated parameters.

3 Design aspects and criteria

Suppose plans are made to collect data to answer a specific research question. In the design phase of the study, there are often several possible choices regarding how the data could be collected. These choices often affect how well the study answers the posed research question. When the research question is formulated as an estimation problem, the study design may affect the variance and bias of the estimators. The task is to choose the design so that the variance and bias are minimised.

In this thesis, the interest is in the estimation of the rate matrix $\mathbf{q}(\tau)$ when only observed transition frequencies over some time intervals are available. There exists some previous research about optimisation of designs for the estimation of transition rates of certain types of continuous-time Markov processes based on repeated current status data. These include the simple birth process [50], the simple death process [51], and the simple birth-death process [52]. Multistate models have been considered for the analysis of epidemic data [53]. In the above presented work the processes either die out or grow indefinitely and thus do not have the ergodic property. Optimal allocation of measurements for ergodic Markov processes have been considered under the two-state model [27] and under multi-state models to estimate species migration rates [54]. Comparisons of efficiencies of rate estimators using discrete observations with respect to the continuous one have also been presented [27, 55]. Optimisation in these work is based on the Fisher information that depends on the model parameters. Generally, design methodologies aiming to relax the dependency on a single parameter value include sequential designs [56], use of a prior distribution [57, 58] and application of the maximin approach [59].

3.1 Utility-based designs

We approach design problems from a utility theoretic point of view, which is common in Bayesian design [60, 61], and then discuss how these are related to classical approaches in which the Fisher information is employed. Formally, let η denote a candidate design that characterises one of the choices and $\boldsymbol{\Upsilon}$ the space of all possible designs. A candidate design η is evaluated using a utility function $U(\eta, \boldsymbol{\Theta}, \boldsymbol{y})$ that depends on the model parameters $\boldsymbol{\Theta}$ and on data $\boldsymbol{y}$ yet to be collected. The dependence on $\boldsymbol{y}$ is handled by taking the expectation over all possible data:

$$U(\eta, \boldsymbol{\Theta}) = \int_{\boldsymbol{y}} U(\eta, \boldsymbol{\Theta}, \boldsymbol{y}) p(\boldsymbol{y}|\boldsymbol{\Theta}, \eta) d\boldsymbol{y}.$$

With regard to dependence on $\boldsymbol{\Theta}$, a common choice is to integrate over a selected prior distribution:

$$U(\eta) = \int_{\Theta} U(\eta, \boldsymbol{\Theta}) p(\boldsymbol{\Theta}) d\boldsymbol{\Theta}.$$

A special case of a prior is the degenerate distribution: $\mathbb{P}\{\boldsymbol{\Theta} = \boldsymbol{\Theta}_0\} = 1$, i.e., a single value of the parameter vector ($\boldsymbol{\Theta}_0$). The maximiser of the utility under a degenerate prior is referred to as the locally optimal design.

The general optimisation task is to maximise the expected utility $U(\eta)$ in a design space $\eta \in \boldsymbol{\Upsilon}$. A common approach in measuring the performance of design η, when concerned with making inferences about a parameter vector $\boldsymbol{\Theta}$, is to choose the utility function so that it involves some properties of an estimator $\hat{\boldsymbol{\Theta}}(\boldsymbol{y})$. A natural requirement is that error of the estimator is minimised. Denoting the parameter vector as $\boldsymbol{\Theta} = (\Theta_1, ..., \Theta_d)$, a possible utility function aiming to meet this criterion is

$$\begin{aligned} U(\eta) &= -\int_{\Theta}\int_{\boldsymbol{y}} (\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}(\boldsymbol{y}))^T \mathbf{W} (\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}(\boldsymbol{y})) p(\boldsymbol{y}|\boldsymbol{\Theta}, \eta) p(\boldsymbol{\Theta}) d\boldsymbol{y} d\boldsymbol{\Theta} \\ &= -\int_{\Theta} \sum_{i=1}^{d} w_i \mathrm{E}[(\Theta_i - \hat{\Theta}_i(\boldsymbol{y}))^2] p(\boldsymbol{\Theta}) d\boldsymbol{\Theta}, \end{aligned} \tag{3}$$

where the expectation is with respect to $p(\boldsymbol{y}|\boldsymbol{\Theta}, \eta)$ and $\boldsymbol{W}$ is a diagonal matrix with elements $w_1, ..., w_d$ with w_i as a subjectively selected importance weight for the component i of $\boldsymbol{\Theta}$. In general, the above utility measures the mean squared error which is the sum of the variance and the squared bias. If the estimator is unbiased then $\mathrm{E}[(\Theta_i - \hat{\Theta}_i)^2] = \mathrm{Var}(\hat{\Theta}_i)$ and the utility measures the weighted average of variances $\mathrm{Var}(\hat{\Theta}_i)$ with weights w_i. Covariances of the parameters ($\mathrm{Cov}(\Theta_i, \Theta_j)$) can be involved in the utility by choosing a non-diagonal $\boldsymbol{W}$. In the locally optimal approach, $p(\boldsymbol{\Theta})$ is replaced by an initial guess of the parameter vector $\boldsymbol{\Theta_0}$, so that the utility function becomes

$$U(\eta; \boldsymbol{\Theta}_0) = -\int_{\boldsymbol{y}} (\boldsymbol{\Theta}_0 - \hat{\boldsymbol{\Theta}}_0(\boldsymbol{y}))^T \boldsymbol{W} (\boldsymbol{\Theta}_0 - \hat{\boldsymbol{\Theta}}_0(\boldsymbol{y})) p(\boldsymbol{y}|\boldsymbol{\Theta}_0, \eta) d\boldsymbol{y}.$$

The evaluation of these utility functions can be based on Monte-Carlo integration. In the locally optimal case, data $\boldsymbol{y}_{c_2}, c_2 = 1, ..., C_2$ are repeatedly sampled from $p(\boldsymbol{y}|\boldsymbol{\Theta}_0)$. For each sampled data set, an estimate $\hat{\boldsymbol{\Theta}}_{\mathbf{0}}(\boldsymbol{y}_{c_2})$ is obtained as the maximum likelihood estimate or the Bayesian posterior mean. The Monte-Carlo approximation of the expected utility is then

$$U(\eta; \boldsymbol{\Theta}_0) \approx -\frac{1}{C_2} \sum_{c_2=1}^{C_2} (\boldsymbol{\Theta}_0 - \hat{\boldsymbol{\Theta}}_0(\boldsymbol{y}_{c_2}))^T \boldsymbol{W} (\boldsymbol{\Theta}_0 - \hat{\boldsymbol{\Theta}}_0(\boldsymbol{y}_{c_2})).$$

If a proper prior distribution $p(\boldsymbol{\Theta})$ is used, the only difference is that the parameter vector is sampled from the prior instead of using a fixed value. The evaluation of (3) can then be based on:

$$U(\eta) \approx -\frac{1}{C_1 C_2} \sum_{c_1=1}^{C_1} \sum_{c_2=1}^{C_2} (\boldsymbol{\Theta}_{c_1} - \hat{\boldsymbol{\Theta}}_{c_1}(\boldsymbol{y}_{c_2}))^T W (\boldsymbol{\Theta}_{c_1} - \hat{\boldsymbol{\Theta}}_{c_1}(\boldsymbol{y}_{c_2})),$$

where $\boldsymbol{\Theta}_{c_1} \sim p(\boldsymbol{\Theta})$ and $\boldsymbol{y}_{c_2} \sim p(\boldsymbol{y}|\boldsymbol{\Theta}_{c_1}), c_2 = 1, ..., C_2, c_1 = 1, ..., C_1$.

In maximum likelihood estimation, the evaluation of variances is often based on the Fisher information matrix $\mathcal{I}(\boldsymbol{\Theta}) = (\mathcal{I}(\boldsymbol{\Theta})_{i,j})$, where

$$\mathcal{I}(\boldsymbol{\Theta})_{i,j} = \mathrm{E}\left[\left(\frac{\partial}{\partial\Theta_i}\log L(\boldsymbol{\Theta};\boldsymbol{y})\right)\left(\frac{\partial}{\partial\Theta_j}\log L(\boldsymbol{\Theta};\boldsymbol{y})\right)\right]$$

and the expectation is taken over data $\boldsymbol{y}$. The inverse of the Fisher information is the variance matrix of the asymptotically normally distributed maximum likelihood estimators ($\mathrm{Var}(\hat{\boldsymbol{\Theta}}) = \mathcal{I}(\boldsymbol{\Theta})^{-1}$). Optimisation of designs when using the Fisher information is commonly based on the A-, D-, or E-optimality criteria, which seek to minimise the trace of $\mathcal{I}(\boldsymbol{\Theta})^{-1}$, the determinant of $\mathcal{I}(\boldsymbol{\Theta})$, or maximise the minimum eigenvalue of $\mathcal{I}(\boldsymbol{\Theta})$, respectively. The D-optimality criterion can be interpreted as the volume of the variance ellipsoid of the parameter estimators, and the E-optimality minimises the maximum variance of normalised linear combinations of the parameter estimators. The A-optimality criterion minimises the average variance of the model parameters. By choosing $\boldsymbol{W}$ as the diagonal matrix in (3), the utility corresponds to the trace of $\mathrm{Var}(\hat{\boldsymbol{\Theta}})$ for unbiased estimators and thus asymptotically corresponds to the A-optimality criterion [62].

The above utility functions involve measures related to the general performance of the estimation, given the study design. An alternative approach can be considered when the main interest is not in the estimation of a parameter vector $\boldsymbol{\Theta}$ but in making predictions. Utility functions in these cases can be based on some criterion that involves the posterior predictive distribution $p(\boldsymbol{y'}|\boldsymbol{y},\eta)$. Practical limitations related to specific study settings can be considered as part of the utility as well. For instance, if the budget of the study is flexible rather than a strict constraint, sampling cost can be part of the utility function [57]. Another interesting example is the battery life of a GPS transmitter that submits the current status of the process and as such built in to the estimation problem [54].

Two-phase designs. In a two-phase study, data are collected in two phases with the possibility to improve the design for the second phase. The first phase data $\boldsymbol{Y}^0(\eta_0) = \boldsymbol{Y}^0$ are collected using an initial design η_0, which can be derived as described above. The question for the second phase is to optimise design η_1 employing the already accumulated information $\boldsymbol{Y}^0$.

Let $\boldsymbol{y}$ denote the second phase observations which are to be collected using the second-phase design η_1. For the optimisation of the second phase design, the utility function is

$$\tilde{U}(\eta_1;\boldsymbol{Y}^0) = -\int_{\Theta}\int_{\boldsymbol{y}}(\boldsymbol{\Theta}-\hat{\boldsymbol{\Theta}}(\boldsymbol{y},\boldsymbol{Y}^0))^T W(\boldsymbol{\Theta}-\hat{\boldsymbol{\Theta}}(\boldsymbol{y},\boldsymbol{Y}^0))^T p(\boldsymbol{y}|\boldsymbol{\Theta},\eta)p(\boldsymbol{\Theta}|\boldsymbol{Y}^0)d\boldsymbol{y}d\boldsymbol{\Theta}. \quad (4)$$

The above utility function $\tilde{U}(\eta_1;\boldsymbol{Y}^0)$ can be used in the optimisation of the second phase design but not in answering questions such as whether the study should be conducted in one or two phases. This is because data $\boldsymbol{Y}^0$ in the utility above are already observed whereas the decision about whether conducting the study in one or two phases needs to be done before collecting any data. The question of whether the study should be conducted in one or two phases needs to be answered by comparing a utility $U(\eta_0)$ of the form (3), in which all data are collected using design η_0, to a two-phase utility where η_0 is the same design for the first phase and η_1 is the second phase design that maximises (4). Because the first phase data are not observed they need to be integrated out. Design η_1 also depends on the first phase data, so $U(\eta_0)$ is compared to the utility

$$\tilde{U}(\eta_0) = \int_{\boldsymbol{Y}^0}\tilde{U}(\eta_1;\boldsymbol{Y}^0)p(\boldsymbol{Y}^0)d\boldsymbol{Y}^0.$$

The distribution $p(\boldsymbol{Y}^0)$ can be specified e.g. to follow a model with a single value of the parameter vector $p(\boldsymbol{Y}^0|\boldsymbol{\Theta}_0)$.

The above presented utility-theoretic approach for the two-phase studies is analogous with the following approach in which the variance of the model parameters is based on the Fisher information. The Fisher information for the first phase is replaced by the observed information, i.e., the sample-based information matrix. It can be calculated as the negative Hessian matrix of the logarithm of the likelihood given the observed data, denoted $-H(\boldsymbol{\Theta_0}; \boldsymbol{Y}^0)$. Based on $\boldsymbol{Y}^0$, a maximum likelihood estimator $\hat{\boldsymbol{\Theta}}_0$ is obtained. The information used to design the second phase with a fixed $\boldsymbol{Y}^0$ is

$$-H(\hat{\boldsymbol{\Theta}}_0; \boldsymbol{Y}^0) \; + \; I(\hat{\boldsymbol{\Theta}}_0; \boldsymbol{Y}^1|\boldsymbol{Y}^0),$$

where the conditioning $\boldsymbol{Y}^1|\boldsymbol{Y}^0$ indicates the possibility of a Markovian dependence of the future data on $\boldsymbol{Y}^0$.

When considering questions such as whether the study should be conducted in one or two phases, the first phase data $\boldsymbol{Y}^0$ need again be integrated out. The information of the two-phase study is then

$$E_{\boldsymbol{Y}^0}[-H(\hat{\boldsymbol{\Theta}}_0; \boldsymbol{Y}^0) + I(\hat{\boldsymbol{\Theta}}_0; \boldsymbol{Y}^1|\boldsymbol{Y}^0)],$$

The method including the observed information has previously been employed by Karvanen, Kulathinal, and Gasbarra [63].

3.2 Design of pneumococcal colonisation studies

This section considers practical issues in designing pneumococcal colonisation studies in which the aim is to estimate the dynamics of colonisation (i.e. the rate matrix). A common framework in such studies is one in which the state of individual i ($i = 1, ..., N$) is measured K_i times. The total number of collected samples is $K^{\text{tot}} = \sum_i K_i$. The distribution of the first observation $\pi(Y_i(\tau_i^1))$, needs to be specified. The subsequent measurements are taken at times $\tau_i^2, ..., \tau_i^{K_i}$ and the distribution of samples $Y_i(\tau_i^j), j = 2, ..., K_i$, follows from the transition probability matrix generated by the rate matrix $\mathbf{q}(\tau)$ parametrised through a vector $\boldsymbol{\Theta}$. The duration of the study is $T = \max_{i,j}\{\tau_i^j\} - \min_i\{\tau_i^1\}$. The design questions in this thesis are related to estimation of $\boldsymbol{\Theta}$ with minimum variance. To achieve this goal, one can consider the choice of the number of study subjects (N), the number of repeated samples per individual (K_i), and the placement of time points $\tau_i^j, i = 1, ..., N, j = 1, ..., K_i$.

The total number of samples (K^{tot}), determined by the number of individuals N and the number of repeated samples per individual $(K_1, ..., K_N)$ is in many cases fixed in advance. For instance, in pneumococcal applications considered in the articles of this thesis, the total number of samples could have been limited by the study budget. K^{tot} is also related to the absolute level of variance that can be attained for the parameter estimators ($\text{Var}(\hat{\boldsymbol{\Theta}})$). In this thesis, K^{tot} is not subject to optimisation and is usually chosen to be "large." The reason for this choice is the aim to eliminate the role of the bias of the estimators in the utility function. The design task under this setting is to minimise the variance of the estimators over the design space.

Even with a fixed total size, the number of samples can be split in various ways across individuals N and repeated measurements per individual K_i. The first sample ($Y_i(\tau_i^1)$) may have a different distribution as compared to the subsequent ones ($Y_i(\tau_i^j)$).

Added information from a new individual may then be different from that gained adding repeated observations for the current study subjects. With regard to K_i, unequal numbers of repeated samples across individuals may be reasonable if, for instance, individuals are known to differ from each other. In most cases, we assume that individuals are identical and consider equal numbers of samples across each individual ($K_i = K$). In Article III, the choice between N and K, conditioned on K^{tot}, is considered for the two-state model.

The stationary distribution is a common choice to model the first observation $\pi(Y_i(\tau_i^1))$. In many pneumococcal colonisation studies this assumption is realistic or not of major concern. Although the prevalence of colonisation is practically zero at birth, it levels off in few months after birth and typically remains relatively stable up to several years of age [64]. For infant studies, the steady-state assumption does not necessarily pose a major concern if most samples are not taken during the first months of life. If the focus is on the estimation of transition rates soon after birth, i.e., until stationarity is attained, Article III shows that at least in models with two states optimal equidistant time spacings do not depend considerably on the initial distribution. However, there may be better performing, non-equidistant designs. For instance, if the non-colonised state is rare under the stationary distribution and also the initial state for most individuals, optimal time spacings can be expected to be shorter during the transition phase.

Sometimes the researcher is in a position to influence the initial distribution. An example are challenge studies in which a proportion of subjects are exposed to a certain condition. It is also possible that the time of first observation is deliberately chosen in order to have indirect influence on the initial distribution. For example, infants may be less prone to infections than older children that have been exposed to the infection pressure for a longer time. In Article III that considers two-state models, the optimal initial distribution is determined and the dependency of optimal time spacings on the initial distribution is investigated. In Articles IV and V the stationary initial distribution is assumed. We note that although the stationary distribution is realistic in many applications it is not necessarily the optimal one.

The allocation of time points for observations $(\tau_i^1, ..., \tau_N^{K_N})$ is the primary design question in Articles III, IV and V. There might be some limitations in how these can be chosen in practice. One possible limitation in allocating the time points is posed by the duration of the study T. For instance, a very long study may raise a concern since the subjects may be lost from the follow-up. Another matter is that the study protocol may not allow scheduled samples to be collected at arbitrary time points, but rather in certain time windows. Although not necessary, for the processes considered in this thesis, sampling times that are within a small time window can be considered as observed on the same day. In this thesis, we do not restrict the duration of the study, but limit to studying designs in which all individuals are sampled at the same time, i.e., $\tau_1^j = ... = \tau_N^j = \tau^j, j = 1, ..., K$.

Another constraint on the design space regarding the selection of times $\tau^1, ..., \tau^K$ is that investigations are limited to equidistant spacings. The question is then to optimise the time spacing t between any two consecutive observations. Because constant transition rates are assumed and the usual scenarios consider the stationary initial distribution, the restriction to equidistant designs is not a further limitation. A heuristic reasoning for this is that, with the stationary initial distribution and constant rates, each repeated sample has the same optimal time spacing when optimised separately. Thus, it is obvious that when all samples are optimised together, they still have the same optimal time spacing. Based on the numerical analyses by Albert and Brown [27], this also holds true for models with two states and two constant rates. When the rates are not constant or

the initial distribution the stationary one, equidistant designs are not necessarily optimal. To optimise the time spacings between consecutive observations, it is then necessary to go through all possible allocations of possible repeated samples. This forms a huge design space if the number of repeated samples per individual is not very small. In practice, it may be possible to investigate only certain types of allocation schemes based on intuitive reasoning.

4 Computational details

Because of the need to integrate over missing data, inference on the model parameters in Articles I, II, and V could not be performed using standard packages in statistical software such as R or Matlab®. In these articles, Markov Chain Monte Carlo methods with the Metropolis-Hastings algorithm were employed to draw samples from the posterior distribution of the model parameters. In Articles III and IV, obtaining maximum likelihood estimates was straightforward, but the computational burden came from the minimisation of the variances of the estimators in a large space of candidate designs. This section highlights some computational aspects of the statistical methods used in this thesis.

In Articles II − V, data were missing by design and the unobserved part of the process between any two consecutive observations was integrated with the use of the matrix logarithm function. This was possible due to the assumption of time-homogeneous transition rates. In Article I, such missing data were handled using an alternate approach. In particular, statistical inferences were based on the continuous-time likelihood (1) after constructing fixed episodes of being susceptible or colonised by one of the serotypes in the model. The construction of episodes was based on discrete-time data, assuming that the process makes a minimum number of state changes between any two consecutive observations. In practice, this meant that at most one unobserved state was needed to be augmented in between any two consecutive observations. The augmented state and the change point where one episode ends and a new one starts were treated as random variables.

Article II involved another type of problem in which the missing data distribution depends on the underlying unobserved state. In this study, it was acknowledged that two serotypes simultaneously colonising the same host may have been detected with less than 100% sensitivity. A recursive algorithm that built all possible transition paths, i.e. considered the possibility that any observation of a singly colonised state might represent a doubly colonised one, was employed. Because it would have been difficult to estimate the sensitivity to observe two simultaneously colonising serotypes, a range of fixed values of the sensitivity was used. This method proved to be computationally rather slow. An alternative method known as the Viterbi algorithm that does not consider all possible transition paths for the underlying hidden state, but only the one with maximum likelihood, would probably have been sufficient.

In Articles III, IV and V, the minimisation of the variance of the rate estimators was of particular interest. In Article I, the variance was based on the Fisher information, evaluated numerically in Matlab® using an analytical but lengthy expression that was derived using symbolic calculation in Maple®. Due to the large number of parameters in the models of Article IV, the Fisher information was not be derived analytically. Instead, data were sampled repeatedly using a model with the true parameter values. For each iteration of data, the maximum likelihood estimates were derived. The measure of uncertainty in the estimates was evaluated as the average squared distance of the estimates

from their true values. In Article V, posterior distributions of the model parameters were sampled using the Metropolis-Hastings algorithm. The measure of uncertainty was based on quantiles of these samples.

Article V considered a hierarchical Bayesian model to estimate vaccine efficacy allowing a heterogeneous response to vaccination at the individual level. At the population level, the parameters included the absolute and relative transition rates that were assumed to be equal and constant across all individuals. Individual-level parameters were needed to account for possible differences across individuals in their response to vaccination. There are multiple ways to implement the same model. We employed an explicit random effect that was assigned for each individual. This variable followed a specified prior distribution with population level parameters. The population level parameters had their own hyperprior. In the Metropolis-Hastings algorithm that was used, the explicit individual-level random effects were updated at each iteration similarly as were the population level parameters.

5 Summary of the main findings

Markov processes with a finite state space can be used to model countless phenomena in many fields of science. Therefore, researchers often encounter the problem of estimating transition rates governing the dynamics of such processes. In many applications, the estimation of transition rates needs to be based on a set of repeated measurements of the current status of the process. Study designs for collecting such data may have a particularly important role regarding how well these rates can be estimated.

An example of an estimation problem concerning the transition rates of a finite state Markov process is pneumococcal colonisation. The motivation to consider this problem in this thesis came from the PneumoCarr project ($2006 - 2011$), which investigated the possibility of using vaccine efficacy against colonisation as part of the licensure process for pneumococcal vaccines [22]. Already at the time of initiation of the PneumoCarr project there were concerns that serotypes for which the pneumococcal conjugate vaccines provide no protection would eventually replace the diminished colonisation and disease by the vaccine serotypes. New vaccines could thus be useful in the future. The likely reason for such replacement is within-host competition between pneumococcal serotypes or strains, but at the time this phenomenon was poorly understood.

The Markov process approach is suitable for studying competition between pneumococcal strains. In this thesis, the new feature in analysing pneumococcal competition was the definition of competition through states in which two serotypes simultaneously colonise the same host. Conventional sampling methods had not been able to detect more than one serotype per sample, not allowing this type of definition for competition when analysing observed data sets [30, 65]. The research made in this thesis (Article I) provided evidence for strong competition which was found to work in a way that serotype already colonising the human nasopharynx reduces the acquisition rate of other serotypes. This finding can be important for predicting replacement patterns because disease is often a consequence of a newly acquired colonisation episode [66].

Although new data with observations on simultaneous pneumococcal colonisation were available for this thesis, there were concerns that competition is estimated too strong if the sensitivity to detect both simultaneously colonising serotypes is poor but this is not accounted for in the analysis. Methodologically, this type of problem can be dealt with the use of a hidden Markov model. In the application of such models (Article II), time

spacings between consecutive observations were found to have an important role regarding what type of bias would occur in the estimation of competition when assuming too high sensitivity of detecting two simultaneously colonising serotypes. In particular, with short time spacings the estimation of competition that works through reducing acquisition rates was found to be unbiased even when analysing the data under an incorrect assumption of perfect sensitivity to detect double colonisation. The re-analyses of the pneumococcal colonisation data accounting for the possibility to detect simultaneous colonisation with less than 100% sensitivity confirmed earlier findings about strong competition that works through reducing acquisition.

New pneumococcal colonisation studies may be important in the future, and it may be of particular interest to investigate the dynamics of colonisation under the effect of new pneumococcal vaccines. In previous pneumococcal colonisation studies little or no justification for choosing a particular study design has been presented. The most obvious question concerns time spacings between consecutive samples that has varied across studied with no apparent reasoning. These notions acted as the motivation to develop a better design methodology for future colonisation studies.

An approximation to the optimal time spacing to collect repeated samples to estimate transition rates was presented as the reciprocal of the maximal sum of rates inwards and outwards from any state in the model (Articles III and IV). This approximation is applicable when the rates are constant and do not differ markedly from each other. Furthermore, when the rates are constant and the initial distribution is stationary, the same time spacing is applicable for any two consecutive samples in a study (equidistant design). This thesis treats pneumococcal colonisation as an example. However, the underlying model in Articles III and IV is a general continuous-time Markov process and the suggested designs are thereby widely applicable.

An important aim in studies of pneumococcal vaccines and any new vaccine is to predict the impact of vaccination in a population. In these predictions, it is substantial to correctly specify the type of vaccine protection. It is possible that a vaccine gives complete protection against infection or it reduces the instantaneous risk to acquire infection with a certain magnitude. These two types of vaccines may give the same average protection across individuals but when given to a population they may have different impact. If the specification of the type of vaccine protection needs to be determined by estimation using repeated current status data, the time spacing between consecutive samples has an important role. An important finding in this thesis (Article V) was that when the type of vaccine protection needs to be inferred from data, the time spacing should be optimized as if the vaccine would not give complete protection to any individual. The number of repeated samples per individual should then be large enough to assure that, if not completely protected from the infection, one is observed with the infection at least once.

There remains several interesting questions regarding the design of longitudinal studies in which observed discrete-time data arises from an underlying continuous-time stochastic process. For studies in which transmission of infectious diseases is investigated, optimal time spacing between consecutive observations would be useful. In transmission studies, the same time spacing between all samples would likely not be optimal because of time varying exposure. This study addressed rather general aspects of data collection for Markov processes. It would be interesting to compare these results to highly specific study designs that account for a variety of study specific factors.

6 Article summaries

Article I, Between-strain competition in acquisition and clearance of pneumococcal carriage – epidemiologic evidence from a longitudinal study of day-care children

A longitudinal dataset of pneumococcal colonisation in unvaccinated day-care children was analysed to investigate within-host competition between pneumococcal strains. The main research questions were whether colonisation with one serotype decreases the acquisition rate of other serotypes as compared to acquisition when uncolonised and whether concurrent colonisation of two serotypes enhances the clearance of these types as compared to the situation when serotypes colonise the host one at a time.

Statistical inference was based on a likelihood that assumed continuous-time observations. Because the actual observations were discrete-time, episodes of being susceptible or colonised with one or two of the serotypes were first constructed. In this construction, it was assumed that during the unobserved time period between any two consecutive observations, the colonisation process made a minimum number of state changes required for compatibility with the observations.

Ongoing pneumococcal colonisation by any one serotype was found to protect from further acquisitions by other serotypes. Clearance of any single serotype was not affected by concurrent colonisation with other serotypes. To our knowledge, this was the first study providing empirical evidence of between-strain competition defined through concurrent colonisation by two serotypes.

Article II, Competition between Streptococcus pneumoniae strains: implications for vaccine-induced replacement in colonization and disease

The aim of this study was to repeat the analysis of Article I and, with use of hidden Markov models to this novel application, take into account the possibility that imperfect sensitivity to detect multiple simultaneously colonising serotypes may affect the estimates of competition between pneumococcal strains. In addition to the data set already analysed in Article I, two new data sets were employed to investigate between-strain competition in epidemiologically different settings. Simulation studies were used to gain further insight into how inferences may be biased when wrongly assuming that multiple colonisation has been observed with a high sensitivity. Different time spacings for sample collection were considered in the simulation studies.

A hidden Markov model was utilised to allow the possibility that when one serotype was observed, this actually arose from a true hidden state in which two serotypes simultaneously colonised the same host. For the probability to observe two serotypes (detection sensitivity), a range of given values was used. Unlike in Article I, the estimation of transition rates was not based on fixed episodes.

When assuming 100% sensitivity to detect two simultaneously colonising serotypes, competition was found to be strong in each of the three datasets, in agreement with the findings in Article I. Simulation studies showed that competition is estimated too strong if the sensitivity of detecting simultaneous colonisation is assumed too high. However, the finding of strong competition remained in all three datasets even when assuming 50% sensitivity.

Article III, Optimal designs for epidemiologic longitudinal studies with binary outcomes

Design questions for collecting longitudinal data under Markov transition models with two states were considered. The rates were estimated using a sequence of observations made at discrete times. The aim was to minimise the variance of either one or both of the transition rate estimators. Particular design questions included the choice of time interval between consecutive observations, the initial distribution of the study subjects and the choice between the number of subjects versus repeated observations per subject. In sequential designs, an upcoming phase of the study was designed using estimates of parameters obtained from data collected in the previous phases.

The covariance matrix for the estimators was derived as the inverse of the Fisher information matrix. Optimal designs were derived by minimising the trace of the covariance matrix (A-optimal design) or by minimising either of the two diagonal elements of the matrix. In sequential designs, the sample based Fisher information matrix (observed information) was used as the measure of accumulated information from the data that have already been collected.

It was shown that the reciprocal of the sum of the two rates in the model is a good approximation for the optimal time spacing between consecutive measurements. The approximation is similar to what had been presented previously, but its applicability was extended by this study. A completely new result was the finding that the initial distribution is important in studies where relatively few repeated samples per individual are collected and that large studies can benefit greatly from sequential designs. It is important, however, that enough samples are collected in the first phase in order to have a high probability to improve the initial design.

Article IV, Optimal observation times for multi-state Markov models – applications to pneumococcal colonisation studies

The main question was to optimise the time spacing between consecutive samples that arise from a Markov transition model with 2, 3, or 4 states. In addition to considering the estimation of transition rates for one homogeneous group, the estimation of rate ratios based on two groups in a comparative study was of interest. Due to the high dimensionality of the problem, arising from allowing many states in the model, some regularity conditions such as the stationary distribution were assumed. Regarding the fact that designs depended on the model parameters, prior distributions were compared to the use of an initial guess of a single value.

The performance of each design was based on the mean squared distance of a sample of parameter estimates from the true parameter values. For the time spacing that minimises the mean squared distance, an approximation was given as the inverse of the maximal sum of the rates inwards and outwards from any particular state in the model. The approximation is easy to apply in practice and is efficient for models with up to at least 4 states when the rates differ from each other less than 10-fold. This result generalises the approximation presented in Article III to models with multiple states.

Article V, Estimation and interpretation of heterogeneous vaccine efficacy against recurrent infections

The problem of estimating vaccine efficacy based on relative transition rates in continuous-time Markov processes in two groups of individuals (vaccinated and unvaccinated) was considered. In the model, recurrent infections with multiple subtypes were considered and estimation was based on discrete-time data. In a group of vaccinated individuals, one can receive complete, partial, or no protection against the infection. A specific research question was whether proportions of individuals gaining complete, incomplete or no protection can be estimated using discrete-time data. Among those who receive incomplete protection, the average magnitude of protection was also one of the parameters to be estimated.

A hierarchical Bayesian approach was employed. General infection dynamics were described using a multistate Markov model with population-level parameters. Individual-level differences in vaccine response were modelled as random effects. Posterior distributions of the model parameters were sampled using the Metropolis-Hastings algorithm. The performance of each design in the estimation of vaccine efficacy was based on the mean and uncertainty in the samples from the posterior distribution.

The best performing time spacing in the estimation of vaccine efficacy depends on the type of vaccine protection. Long time spacings perform better in the estimation of the proportion of individuals who are completely protected. By comparison, shorter time spacings perform better in the estimation of the average magnitude of incomplete protection. If the estimation of the type of vaccine protection fails, appropriate summary measures may still be estimable.

References

[1] Snow J. On the mode of communication of cholera. *John Churchill, London.* 1855.

[2] Editorials. Looking back on the millennium in medicine. *New England Journal of Medicine.* 2000;**342**:42-9.

[3] Anderson RM, May RM. Infectious diseases of humans: dynamics and control. *Oxford and New York: Oxford University Press.* 1991.

[4] Diekmann O, Heesterbeek JAP. Mathematical epidemiology of infectious diseases: model building, analysis and interpretation. *Wiley.* 2000.

[5] Ross R. The prevention of malaria. *E.P. Dutton & Company.* 1910.

[6] Ross R. The prevention of malaria. *London: John Murray.* 1911;651-86.

[7] Lotka AJ. Quantitative studies in epidemiology. *Nature.* 1912;**88**:497-498.

[8] Lotka AJ. Contributions to the analysis of malaria epidemiology. II General part (continued). Comparison of two formulae given by Sir Ronald Ross. *American Journal of Epidemiology.* 1923;**3**(supp1):38-54.

[9] Kermack WO and McKendrick AG. A Contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A.* 1927;**115**:700-21.

[10] MacDonland G. The analysis of equilibrium in malaria. *Tropical Diseases Bulletin.* 1952;**49**:813-29.

[11] MacDonald G. The measurement of malaria transmission. *Proceeding of the Royal Society of Medicine.* 1955;**48**:295302.

[12] Bailey NTJ. The mathematical theory of infectious diseases and its applications. 2nd ed. *Hafner Press*, New York. 1975.

[13] Becker NG. Analysis of infectious disease data. *Chapman et Hall/CRC.* 1989.

[14] Andersen PK, Borgan O, Gill RD, Keiding N. Statistical models based on counting processes. *Springer.* 1993.

[15] Halloran ME, Longini Jr. IM, Struchiner CJ. Design and analysis of vaccine studies *Springer.* 2010.

[16] Choi YH, Jit M, Flasche S, Gay N, Miller E. Mathematical modelling long-term effects of replacing Prevnar7 with Prevnar13 on invasive pneumococcal diseases in England and Wales. *PLoS ONE.* 2012;**7**(7):e39927.

[17] Nurhonen M, Cheng AC, Auranen K. Pneumococcal transmission and disease in silico: a microsimulation model of the indirect effects of vaccination. *PLoS ONE.* 2013;**8**:e56079.

[18] O'Brien KL, Wolfson LJ, Watt JP, Henkle E, Deloria-Knoll M, McCall N, et al. Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: global estimates. *Lancet.* 2009;**374**:893-902.

[19] Simell B, Auranen K, Käyhty H, Goldblatt D, Dagan R, O'Brien KL. Pneumococcal Carriage Group. The fundamental link between pneumococcal carriage and disease. *Expert Review of Vaccines.* 2012;**11**:841-55.

[20] Eskola J, Kilpi T, Palmu A, Jokinen J, Haapakoski J, Herva E, et al. Efficacy of a pneumococcal conjugate vaccine against acute otitis media. *The New England Journal of Medicine.* 2011;**344**:403-9.

[21] Weinberger DM, Malley R, Lipsitch M. Serotype replacement in disease after pneumococcal vaccination. *Lancet.* 2011;**378**:1962-73.

[22] Goldblatt D, Ramakrishnan M, O'Brien K. Using the impact of pneumococcal vaccines on nasopharyngeal carriage to aid licensing and vaccine implementation; a PneumoCarr meeting report March 27-28, 2012, Geneva. *Vaccine.* 2013;**32**(1):146-52.

[23] Satzke C, Turner P, Virolainen-Julkunen A, Adrian PV, Antonio M, Kim MH, et al. Standard method for detecting upper respiratory carriage of Streptococcus pneumoniae: updated recommendations from the World Health Organization pneumococcal carriage working group. *Vaccine.* 2013;**32**:165-79.

[24] Lipsitch M. Vaccination against colonizing bacteria with multiple serotypes. *Proceeding of the National Academy of Sciences USA.* 1997;**94**(12):6571-6

[25] Van Effelterre T, Moore MR, Fierens F, Whitney CG, White L, Pelton SI, Hausdorff WP. A dynamic model of pneumococcal infection in the United States: implications for prevention through vaccination. *Vaccine.* 2010;**28**:3650-60.

[26] Pradas R, Gil de Miguel A, Álvaro A, Gil-Prieto R, Lorente R, Méndez C, Guijarro P, Antonañzas F. Budget impact analysis of a pneumococcal vaccination programme in the 65-year-old Spanish cohort using a dynamic model. *BMC Infectious Diseases.* 2013;**13**:175.

[27] Albert PS, Brown CH. The design of a panel study under an alternating Poisson process assumption. *Biometrics.* 1991;**47**:921-32.

[28] Siamak PY, Dominique AC. Neisseria meningitidis: an overview of the carriage state. *Journal of Medical Microbiology.* 2004;**53**:821-32.

[29] Hayward AC, Fragaszy EB, Bermingham A, Wang L, Copas A, Edmunds WJ, et al. Comparative community burden and severity of seasonal and pandemic influenza: results of the Flu Watch cohort study. *The Lancet Respiratory Medicine.* 2014;**2**:445-54.

[30] Lipsitch M, Abdullahi O, D'Amour A, Xie W, Weinberger DM, Tchetgen, et al. Estimating rates of carriage acquisition and clearance and competitive ability for pneumococcal serotypes in Kenya with a Markov transition model. *Epidemiology.* 2012;**23**:510-9.

[31] Auranen K, Arjas E, Leino T, Takala AK. Transmission of pneumococcal carriage in families: a latent Markov process model for binary data. *Journal of the American Statistical Association.* 2000;**95**:1044-53.

[32] Cox DR, Miller HD. The theory of stochastic processes. *London: Methuen.* 1965.

[33] Stroock WD. An Introduction to Markov Processes. *Springer* 2004.

[34] Kalbfleisch J, Prentice R. The Statistical Analysis of Failure Time Data. *New York: Wiley.* 1980

[35] Nelson W. Hazard plotting for incomplete failure data. *Journal of Quality Technology.* 1969;**1**:27-52.

[36] Aalen OO. Nonparametric inference for a family of counting processes. *Annals of Statistics.* 1978;**6**(4):701-26.

[37] Aalen OO. Nonparametric estimation of partial transition probabilities in multiple decrement models Annals of Statistics. *Scandinavian Journal of Statistics.* 1978;**6**:534-45.

[38] Aalen OO, Johansen S. An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics.* 1978;**5**:141-50.

[39] Johansen S. The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics.* 1978;**5**:195-9.

[40] Andersen PK, Keiding N. Multi-state models for event history analysis. *Statistical Methods in Medical Research.* 2002;**11**:91-115.

[41] Bladt M, Sørensen M. Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society, Series B.* 2005;**6**:395-410.

[42] Yong C, Jianmin C. On the imbedding problem for three-state time homogeneous Markov chains with coinciding negative eigenvalues. *Journal of Theoretical Probability.* 2011;**24**:928-38.

[43] Antonín Slavík. Product integration, its history and applications. *Matfyzpress*, Prague. 2007.

[44] Higham NJ, Lin L. On p^{th} roots of stochastic matrices. *Linear Algebra and its Applications.* 2011;**435**:44863

[45] Rubin DB. Inference and missing data. *Biometrika.* 1976;**63**:581-92.

[46] Schafer JL, Graham JW. *Psychological Methods.* 2002;**7**:14777

[47] Kenward MG, Lesaffre E, Molenberghs G. An application of maximum likelihood and generalised estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics.* 1994;**50**(4):945-53.

[48] Cappé O, Moulines E, Ryden T. Inference in hidden Markov models. *Springer.* 2007.

[49] Deng K, Sun Y, Mehta PG, Meyn SP. An information-theoretic framework to aggregate a Markov chain, *Proceedings of American Control Conference.* St. Louis, MO, 2009; pp. 731736.

[50] Becker G, Kersting G. Design problems for the pure birth process. *Advances in Applied probability.* 1983;**15**:255-73.

[51] Pagendam DE, Pollett PK. Locally optimal designs for the simple death process. *Journal of Statistical Planning and Inference.* 2010;**140**:3096-105.

[52] Pagendam DE, Pollett PK. Optimal sampling and problematic likelihood in a simple population model. *Environmental Modeling and Assessment.* 2009;**14**:759-67.

[53] Hwang W, Brookmeyer R. Design of panel studies for disease progression with multiple stages. *Lifetime Data Analysis.* 2003;**9**:261-74.

[54] Pagendam D, Ross JV. Optimal use of GPS transmitter for estimating species migration rate. *Ecological Modelling.* 2013;**249**:37-41.

[55] Cook RJ. Information and efficiency consideration in planning studies based on two-state Markov processes. *Journal of Statistical Research.* 2000;**34**:161-78.

[56] Abdelbasit KM, Plackett RL. Experimental design for binary data. *Journal of the American Statistical Association.* 1983;**78**:90-8.

[57] Quintana FA, Müller P. Optimal sampling for repeated binary measurements. *The Canadian Journal of Statistics.* 2004;**32**:73-84.

[58] Cook AR, Gavin GJ, Gilligan CA. Optimal observation times in experimental epidemic processes. *Biometrics* 2008;**64**:860-8.

[59] Tekle FB, Tan FES, Berge MPF. Maximin D-optimal designs for binary longitudinal responses. *Computational Statistics and Data Analysis.* 2008;**52**:5253-62.

[60] Chaloner K, Verdinelli I. Bayesian experimental design: a review. *Statistical Science.* 1995;**10**:273-304

[61] DasGupta A. Review of optimal Bayes designs. Design and analysis of experiments. *Handbook of Statistics.* 1996;**13**:1099148.

[62] Pukelsheim F. Optimal design of experiments. *John Wiley & Sons, Inc.*, New York, 1993.

[63] Karvanen J, Kulathinal S, Gasbarra D. Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates. *Computational Statistics & Data Analysis.* 2009;**53**:1782-93.

[64] Hill PC, Cheung YB, Akisanya A, Sankareh K, Lahai G, Greenwood BM, Adegbola RA. Nasopharyngeal carriage of Streptococcus pneumoniae in Gambian infants: a longitudinal study. *Clinical Infectious Diseases.* 2008;**46**:807-14.

[65] Melegaro A, Choi Y, Pebody R, Gay N. Pneumococcal carriage in United Kingdom families: estimating serotype-specific transmission parameters from longitudinal data. *American Journal of Epidemiology.* 2007;**166**(2):228-35.

[66] Syrjänen RK, Auranen KJ, Leino TM, Kilpi TM, Mäkelä PH. Pneumococcal acute otitis media in relation to pneumococcal nasopharyngeal carriage. *Pediatric Infectious Disease Journal.* 2005;**24**(9):801-6.