Received: 28 February 2021 | Revised: 30 October 2021 | Accepted: 15 December 2021

DOI: 10.1002/sim.9305

RESEARCH ARTICLE

Statistics in Medicine WILEY

Approximate likelihood-based estimation method of multiple-type pathogen interactions: An application to longitudinal pneumococcal carriage data

Irene Man[1,2] | **Johannes A. Bogaards**[1,3] | **Kishan Makwana**[1] | **Krzysztof Trzciński**[4] | **Kari Auranen**[5,6]

[1]Centre for Infectious Diseases Control, National Institute for Public Health and the Environment, Utrecht, The Netherlands

[2]Julius Centre, UMC Utrecht, Utrecht University, Utrecht, The Netherlands

[3]Department of Epidemiology & Data Science, Amsterdam University Medical Centres, Amsterdam, The Netherlands

[4]Department of Pediatric Immunology and Infectious Diseases, Wilhelmina's Children Hospital, University Medical Centre Utrecht, Utrecht, The Netherlands

[5]Department of Mathematics and Statistics, University of Turku, Turku, Finland

[6]Department of Clinical Medicine, University of Turku, Turku, Finland

Correspondence
Irene Man, IARC, 150 Cr Albert Thomas, 69008 Lyon, France.
Email: irene.y.s.man@gmail.com

Funding information
Strategic Programme from the National Institute for Public Health and the Environment (RIVM) of the Netherlands, Grant/Award Number: S/113005/01/PT

Abstract

While the serotypes of *Streptococcus pneumoniae* are known to compete during colonization in human hosts, our knowledge of how competition occurs is still incomplete. New insights of pneumococcal between-type competition could be generated from carriage data obtained by molecular-based detection methods, which record more complete sets of serotypes involved in co-carriage than when detection is done by culture. Here, we develop a Bayesian estimation method for inferring between-type interactions from longitudinal data recording the presence/absence of the types at discrete observation times. It allows inference from data containing co-carriage of two or more serotypes, which is often the case when pneumococcal presence is determined by molecular-based methods. The computational burden posed by the increased number of types detected in co-carriage is addressed by approximating the likelihood under a multi-state model with the likelihood of only those trajectories with minimum number of acquisition and clearance events between observation times. The proposed method's performance was validated on simulated data. The estimates of the interaction parameters of acquisition and clearance were unbiased in settings with short sampling intervals between observation times. With less frequent sampling, the estimates of the interaction parameters became more biased, but their ratio, which summarizes the total interaction, remained unbiased. Confounding due to unobserved heterogeneity in exposure could be corrected by including individual-level random effects. In an application to empirical data about pneumococcal carriage in infants, we found new evidence for between-serotype competition in clearance, although the effect size was small.

KEYWORDS

approximate likelihood, Bayesian inference, co-carriage, longitudinal data, multiple-type interactions, *Streptococcus pneumoniae*

1 | INTRODUCTION

A recurrent theme in research of *Streptococcus pneumoniae* (the pneumococcus) is whether and how pneumococcal serotypes interact with each other.[1-8] As pneumococcal conjugate vaccination has led to serotype replacement, it is evident that between-serotype competition exists.[9] However, the mechanisms by which different serotypes compete remain poorly understood.[10,11]

The way pneumococcal serotypes interact with each other during colonization (carriage) in human hosts has often been studied in longitudinal settings.[2-7] Using multi-state models, it has been possible to infer from longitudinal data whether and to what extent different serotypes compete by inhibiting acquisition or enhancing clearance of colonization of one another. In such models, the possible combinations of types with which a host can be simultaneously colonized define the model's states, whereas the transitions between these states represent events of acquisition and clearance. With these definitions, interactions in acquisition and clearance can be quantified in terms of ratios between appropriate transition rates. Competition in acquisition is quantified by reduced rates of acquiring a new serotype in presence relative to absence of other types, whereas competition in clearance is characterized by higher rates of clearing a serotype in presence relative to absence of other types. Based on estimates of such rate ratios, various studies have concluded that competition in acquisition is likely.[2-6] By contrast, only a few studies have also found evidence for competition in clearance.[4,7]

Nevertheless, previous findings of between-serotype interactions may have been biased by the suboptimal sensitivity of culture-based methods for detecting pneumococcal presence.[6,7] While culture-based methods seldom detect co-carriage of more than two types, it is common for molecular-based methods to discover co-carriage of three or more types. Cross-validation of samples using both methods suggests an underdetection by culture-based methods when multiple types are present.[12-14] Previous studies have suggested that underdetection of co-carriage may have biased estimates of between-type interaction in the direction of stronger competition.[6,7] Molecular-level data may provide a more accurate picture of serotype co-occurrence but have not yet been used for the purpose of estimating between-type interaction.

In principle, inference of between-serotype interactions from molecular-level longitudinal data is not different from analyzing culture-level data. It is still possible to define rate ratios of acquisition and clearance as measures of interactions. However, estimation of these measures becomes more computationally demanding when the data only record the states of the underlying dynamics at discrete observation times. For instance, when observing the host to be in the non-carriage state at one observation time and carrying some serotype(s) at the subsequent observation time, an estimation method that considers the full likelihood based on the observed data must account for the many possible time points at which acquisition(s) could have taken place. In addition, it has to take into account the possibility of one or more serotypes being acquired and cleared successively for multiple times between the observations times. Exploration of all possible trajectories of acquisition and clearance compatible with the observed data is hence computationally intensive. To alleviate this computationally intensive task, previously developed estimation methods have restricted the model state space to co-carriage states with up to two serotypes. Such simplification was justified in analyses of culture-level data, as observation of co-carriage with more serotypes was rare. To analyze of molecular-level data with many observations of co-carriage with two or more serotypes, this restriction of state space is no longer justifiable, and new approaches need to be developed to handle such data.

In this article, we develop an estimation method that allows inference of interactions between pathogen types from longitudinal data containing co-carriage of two or more types, as is often the case when common multi-strain pathogens (eg, *S pneumoniae*, human papillomavirus, *Plasmodium falciparum*) are detected with molecular-based methods. The new method tackles the computational task by restricting the likelihood function under a multi-state model to account the likelihood of only those trajectories with a minimal number of transitions. The loss in accuracy resulting from this likelihood approximation likely depends on sampling strategy and requires case-specific consideration. The article is organized as follows. Section 2 specifies the model for multiple-type pneumococcal colonization dynamics. In Section 3, we present an approximation of the likelihood function and embed it in a Bayesian estimation framework. In Section 4, we validate the proposed Bayesian approximate likelihood-based (BALB) estimation method on simulated data. In light of the motivating pneumococcal data set, we investigate various aspects of study design that may affect the accuracy of estimation. In Section 5, we apply the new method to a molecular-level data set of pneumococcal carriage in infants to investigate between-type competition of pneumococcus. Finally, we conclude and discuss our findings in Section 6.

2 | A MULTI-STATE MODEL OF MULTIPLE-TYPE PNEUMOCOCCAL COLONIZATION DYNAMICS

We here specify a continuous-time multi-state model for the pneumococcal colonization dynamics of an arbitrary number of serotypes at the individual level. We assume a Markov model so that the transition rates are constant over time. For the pneumococcal data set we consider in this work, the Markov assumption is justifiable as it consists of infants in their first 18 months of life, during which the build-up of naturally acquired serotype-specific immunity as well as cross-protection across types are limited.[15,16] For simplicity, we also assume any seasonal effects to be negligible.

The model's state space is the set of possible combinations of a given number p of serotypes. Including the uncolonized state $\emptyset$, there are maximally 2^p states possible. Models previously used to infer between-type interactions from culture-based data contained no co-carriage states of more than two types and hence consisted maximally of $1 + p(p + 1)/2$ states.[2-7] In this work, we allow co-carriage states with more than two types, in agreement with what was observed in the pneumococcal data set of this study. In addition, some co-carriage states with more than two types that could occur between the observed co-carriage states are included to the model state space, which will be further specified in Section 3.

Except for allowing for more co-carriage states, other model assumptions (with regard to acquisition, clearance and interaction between serotypes) are similar to previously developed multi-state models of pneumococcal colonization dynamics.[2,4,5,7] Transitions between the model states take place when individual serotypes are acquired or cleared. We only allow acquisition and clearance of one type at a time so that the model accommodates only transitions between those pairs of states that differ by one type (Figure 1). This does not impose a real restriction, as the probability of multiple transitions occurring simultaneously is arguably negligible.

An uncolonized individual acquires type j at rate $z_i\lambda_j$, where λ_j is the per capita baseline acquisition rate of type j and z_i an individual-level parameter ("random effect") indicating the level of exposure and/or predisposition to pneumococcal colonization of individual i. An individual already colonized with one or more of the other types acquires type j at rate $kz_i\lambda_j$. Hence, parameter k is the rate ratio for acquiring a type in presence vs absence of any other types and describes the between-type interaction in acquisition, with $k < 1$ denoting competition.

In a singly colonized individual, that is, in absence of any other types, clearance of type j occurs at a type-specific baseline rate μ_j, whereas clearance in presence of other types occurs at rate $h\mu_j$. Hence, parameter h is the rate ratio for clearing a type in presence vs absence of other types, describing the between-type interaction in clearance, with $h > 1$ denoting competition.

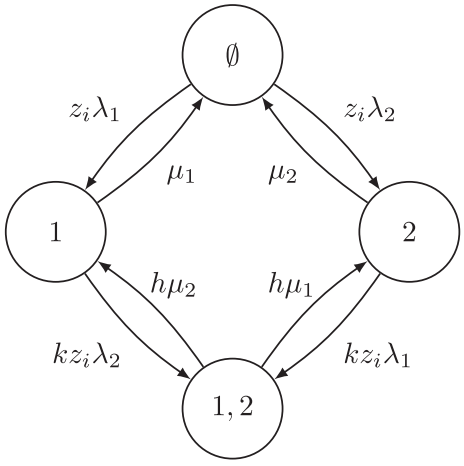

FIGURE 1 Structure of the multi-state model. The depicted multi-state model describes the colonization dynamics of $p = 2$ pneumococcal serotypes at the individual level. This model has $2^p = 4$ states and $2^p \cdot p = 8$ transitions. Shown are the baseline acquisition rates, λ_1 and λ_2, the baseline clearance rates, μ_1 and μ_2, the random effect of individual i, z_i, and the interaction parameters of acquisition and clearance, k and h

In summary, for a given individual i, the transition rate between any two model states, from x to y, is given by

$$Q_i(x,y) = \begin{cases} z_i\lambda_j & \text{if } x = \emptyset \text{ and } y = \{j\}, \\ kz_i\lambda_j & \text{if } x \neq \emptyset, y = x \cup \{j\} \text{ and } j \notin x, \\ \mu_j & \text{if } y = \emptyset \text{ and } x = \{j\}, \\ h\mu_j & \text{if } y \neq \emptyset, x = y \cup \{j\} \text{ and } j \notin y, \\ 0 & \text{if otherwise,} \end{cases} \tag{1}$$

where $\emptyset$ denotes the uncolonized state, $\notin$ not contained by, and $\cup$ the union of the respective sets.

Arranging all transition rates of individual i in a matrix yields a transition rate matrix Q_i, in which the (x, y) element is the transition rate from state x to state y. By convention, each diagonal element of Q_i is specified as the additive inverse of the total transition rate out of the corresponding state, that is, $Q_i(x) = -\sum_{y \neq x} Q_i(x,y)$.

3 | A BALB ESTIMATION METHOD

In this section, we present a BALB method for estimating parameters of the multi-state model of Section 2, when carriage states data D have been recorded from each study subject at a number of discrete observation times. Denote the vector of unknown parameters by $\theta = (\lambda_1, \ldots, \lambda_p, \mu_1, \ldots, \mu_p, k, h, z_1, \ldots, z_m)^T$. For m study subjects, the parameters include $2p$ baseline rates, 2 interaction parameters, and m random effects. The estimation method relies on an approximation of the likelihood function $p(D|\theta)$ of the model parameters θ.

Before writing the likelihood function based on the entire data set D, we specify how the contribution from one consecutive pair of observations from one individual is approximated. For a moment, for notational convenience, we use D to denote this single pair of observations and Q to denote the transition rate matrix of the individual in question.

Suppose that we observe the individual to be in state x_{start} at time t_{start} and in state x_{end} at the subsequent observation time t_{end}, that is, $D = \{x_{\text{start}}, x_{\text{end}}, t_{\text{start}}, t_{\text{end}}\}$. The likelihood contribution $p(D|\theta)$ is given by the total probability of all possible trajectories compatible with this observation, that is, of all paths of connecting states starting in state x_{start} and ending in state x_{end}, with corresponding sojourn times summing up to $\Delta T = t_{\text{end}} - t_{\text{start}}$. Theoretically, this probability can be evaluated by first determining the transition probability matrix $\exp(Q \cdot \Delta T)$ and then taking its $(x_{\text{start}}, x_{\text{end}})$ element. However, calculating the exact transition probability matrix amounts to computing a matrix exponential, which is computationally intensive when the dimension of Q is high, that is, when many states are included in the model.[17] This is the case when many co-carriage states with more than two types are observed. Also any data augmentation approach would be computationally demanding if all possible trajectories need to be simulated. To make the computation of the likelihood computationally feasible, we approximate $p(D|\theta)$ by considering only the subset of all possible trajectories that are compatible with the observation. Specifically, we only consider those trajectories that contain the minimum number of transitions connecting state x_{start} to state x_{end}. We thus include only those trajectories that contain exactly one transition for each type that is in x_{start} but not x_{end}, or vice versa, and disregard any trajectory that involves successive acquisition and clearance of the same type between observation times. In effect, when states x_{start} and x_{end} differ with regard to the status of n types, a minimum-transition trajectory contains exactly $n + 1$ states and n transitions. Note that, due to this restriction to minimum-transition trajectories, the model state space is reduced to only those co-carriage states that show up in a minimum-transition trajectory.

To enumerate all minimum-transition trajectories, let $\mathcal{X}(D)$ and $\mathcal{T}(D)$ denote the collection of all minimum-transition paths and the collection of all compatible sojourn times, respectively. As an example, suppose that $D = \{x_{\text{start}}, x_{\text{end}}, t_{\text{start}}, t_{\text{end}}\} = \{\emptyset, \{1,2\}, 0, 2\}$. The corresponding collection of paths $\mathcal{X}(D)$ consists of two paths: $(\emptyset, \{1\}, \{1,2\})$ and $(\emptyset, \{2\}, \{1,2\})$, while the corresponding collection of sojourn times $\mathcal{T}(D)$ consists of all positive triplets summing up to ΔT, for example, $(0.5, 0.5, 1.0)$ (Figure 2). The approximation of the likelihood contribution $\tilde{p}(D|\theta)$ is given by the likelihood of all minimum-transition trajectories, obtained by enumerating all combinations of paths and sojourn times in $\mathcal{X}(D)$ and $\mathcal{T}(D)$:

$$\tilde{p}(D|\theta) = \sum_{\boldsymbol{x} \in \mathcal{X}(D)} \int_{\boldsymbol{t} \in \mathcal{T}(D)} p(\boldsymbol{x}, \boldsymbol{t}|\theta) d\boldsymbol{t}$$

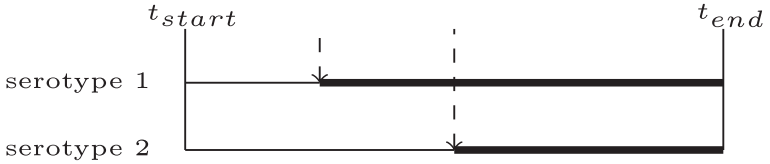

FIGURE 2 Example of a minimum-transition trajectory. The depicted trajectory is one possible minimum-transition trajectory corresponding to data $D = \{x_{\text{start}}, x_{\text{end}}, t_{\text{start}}, t_{\text{end}}\} = \{\emptyset, \{1,2\}, 0, 2\}$. The depicted path is $(\emptyset, \{1\}, \{1,2\})$, with sojourn times $(0.5, 0.5, 1.0)$. Legend: Solid vertical lines denote observation times; dashed vertical arrows denote transition times; thin horizontal lines denote non-carriage; thick horizontal lines denote carriage

$$= \sum_{\boldsymbol{x}\in\mathcal{X}(D)} \int_0^{\Delta T} \int_0^{\Delta T - t_1} \cdots \int_0^{\Delta T - \sum_{r=1}^{n-1} t_r} p(\boldsymbol{x}, t_1, t_2, \ldots, t_n, \Delta T - \sum_{r=1}^{n} t_r | \theta) dt_n \cdots dt_2 dt_1$$

$$= \sum_{\boldsymbol{x}\in\mathcal{X}(D)} \int_0^{\Delta T} \int_0^{\Delta T - t_1} \cdots \int_0^{\Delta T - \sum_{r=1}^{n-1} t_r} e^{Q(x_{n+1})(\Delta T - \sum_{r=1}^{n} t_r)} \prod_{r=1}^{n} e^{Q(x_r)t_r} Q(x_r, x_{r+1}) dt_n \cdots dt_2 dt_1. \quad (2)$$

The integrand on the last line is the standard likelihood in multi-state models for a single trajectory, in which the exponential terms account for sojourning in the visited states and the off-diagonal terms of matrix Q for the respective transitions in between.[18]

Equation (2) can be simplified into the following expression, which is less computationally intense to evaluate (see Supplementary Appendix A for verification):

$$\tilde{p}(D|\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}(D)} e^{Q(x_{n+1})\Delta T} \left(\prod_{r=1}^{n} Q(x_r, x_{r+1}) \right) \left(\frac{(-1)^n}{\prod_{r=1}^{n-1} Q(x_r) - Q(x_n)} + \sum_{r=1}^{n} \frac{e^{(Q(x_r) - Q(x_{n+1}))\Delta T}}{(Q(x_r) - Q(x_{n+1})) \prod_{s=1, s\neq r}^{n} (Q(x_r) - Q(x_s))} \right). \quad (3)$$

See Supplementary Appendix B for an analysis demonstrating the goodness of the approximation under reasonable length of the sampling interval ΔT.

We now return to the general case in which D consists of an arbitrary number of individuals and observation times. An approximation to the entire likelihood is obtained by multiplying the approximate likelihood contributions of all pairs of consecutive observation times, that is,

$$\tilde{p}(D|\theta) = \prod_{i=1}^{m} \prod_{l=1}^{u_i - 1} \tilde{p}(D_{il}|\theta), \quad (4)$$

where u_i is the number of observation times of individual i, and $D_{il} = \{x_l^i, x_{l+1}^i, t_l^i, t_{l+1}^i\}$ the lth consecutive pair of observations of individual i.

Finally, the approximate likelihood function in Equation (4) is embedded into a Bayesian framework, in which statistical inference is enabled by estimating the posterior probability of the model parameters according to Bayes' theorem. The posterior probability $p(\theta|D)$, which is proportional to the product of the likelihood function $p(D|\theta)$ and the prior $p(\theta)$, is approximated through the use of the approximate likelihood $\tilde{p}(D|\theta)$. The proposed estimation method was implemented in the statistical software STAN (https://github.com/irene-man/), which performs Markov chain Monte Carlo simulation with a Hamiltonian Monte Carlo scheme.

4 | SIMULATION STUDY

4.1 | Methods to compare

The BALB estimation method was validated on simulated data. As a benchmark, we compared its performance against a naive method which imputes transition times midway between the consecutive observation times. With the imputed

transition times, the naive method obtains the maximum likelihood estimates of the model parameters by means of Poisson regression (see Supplementary Appendix C for details of the naive method). Note that also the naive method makes the assumption of minimum transitions.

In addition, we investigated how well BALB was able to adjust for bias due to unobserved heterogeneity in exposure and/or predisposition to pneumococcal colonization. To do so, we considered two implementations of BALB, one with individual-specific random effects (z_i) and one without.

4.2 | Simulated settings

We simulated longitudinal data sets according to the multi-state model of Section 2. Each data set consisted of 500 individuals with a follow-up of 20 months. The large number of individuals was chosen to ensure stable estimates, facilitating identification of biases in the estimated parameter values. For each individual, the initial state at time 0 was non-carriage. Data were simulated assuming the same baseline rates across all serotypes: $\lambda_j = \exp(-3.5) \approx 0.030$ per month (acquisition) and $\mu_j = \exp(-1.5) \approx 0.22$ per month (clearance). Roughly, in absence of interactions, these rates would yield a prevalence of 12% for each serotype in the steady state. Parameters that were varied across simulated settings included the number of types (p), the length of the sampling interval (ΔT), the interaction parameters (k and h), and the variance of the individual-level random effects z_i (α).

4.2.1 | BALB and naive methods in absence of heterogeneity in exposure

We compared the performance of BALB and the naive method in various settings in which there was no heterogeneity in exposure. First, setting the number of types to 2, we varied the length of the sampling interval ($\Delta T = 0.5, 1, 2$ months). Then, fixing the sampling interval to 2 months, we varied the number of types ($p = 2, 4, 7$). For each combination of sampling interval and number of types, we explored nine distinct pairs of parameter values for the two interaction parameters, each chosen from (0.5, 1.0, 2.0). For the interaction parameter of acquisition k, these values correspond to competition, independence and synergy, respectively, whereas for the interaction parameter of clearance h, they correspond to synergy, independence and competition, respectively.

4.2.2 | BALB under heterogeneity in exposure

We next investigated whether BALB was able to adjust for unobserved heterogeneity in exposure by including random effects. Heterogeneity was realized by simulating individual-specific random effects z_i using a gamma distribution with mean one and variance α. We simulated settings with different values of α, that is, $\alpha = 0.00001, 0.1, 0.2, 0.5, 1$. In the case with the variance as large as 1, the 20% of individuals with the largest random effects have at least 7.2 times higher rates for type-specific acquisitions than the 20% of individuals with the smallest random effects. At the other extreme, $\alpha = 0.00001$ resembles the homogeneous case. All random effects z_i and α needed to be estimated from the data in order to mimic the situation in which there is no external information (or incomplete information) on variation in exposure levels or susceptibility to colonization across study participants. Throughout, the number of types was fixed at 2 and the length of sampling interval fixed at 1 month. These settings of different values of α were repeated for two combinations of interaction parameters: (1) competition in acquisition and no competition in clearance ($k = 0.5$, $h = 1$) and (2) no competition in acquisition and competition in clearance ($k = 1$, $h = 2$).

4.3 | Prior distributions

We assumed the following prior distributions for the model parameters in BALB. The baseline acquisition and clearance rates, λ_j and μ_j, were assumed to have gamma-distributed priors. The corresponding rate parameters were chosen as the crude baseline rates, which were derived from the data (see Supplementary Appendix D for the derivation). The corresponding shape parameters were fixed to a small value (0.00001) in order to be non-informative. For the interaction parameters, we assumed uniform priors on the log scale with a symmetrical range around zero to allow as much competitive as synergistic interactions, that is, $\log(k), \log(h) \sim \text{Unif}(-3, 3)$. The random effects were assumed to follow a gamma

distribution with mean one and variance α to ensure identifiability, that is, $z_i \sim \text{Gamma}\left(\frac{1}{\alpha}, \frac{1}{\alpha}\right)$, where $\frac{1}{\alpha}$ is both the shape and rate parameter. For $1/\alpha$, we assumed the following log-normal-distributed hyperprior: $1/\alpha \sim \text{Lognormal}(0, 2)$.

4.4 | Performance measures

To test the performance of the proposed method, for each setting we simulated 100 data sets, from which the model parameters were estimated. For the Bayesian methods, estimates are given as posterior means and 95% credible intervals (CIs). For the naive method, estimates are given as maximum likelihood estimates and 95% confidence intervals (also abbreviated as CIs). To assess bias in the estimation of the baseline rates and the log-transformed interaction parameters, we computed the difference between the true parameter values and the mean of the estimates across the 100 simulated data sets. In addition, we also assessed the bias in the estimates of the log-transformed ratio of the two interaction parameters, $\log(k/h)$.

4.5 | Results

4.5.1 | Performance of BALB and the naive methods in absence of heterogeneity in exposure

With 2 types and a short sampling interval of 0.5 months, both the BALB and naive methods were able to estimate the interaction parameters with little bias (Figure 3A, eTable 2 in Supplementary Appendix E). With increasing length of the sampling interval, both methods led to biased estimation, except when there was no interaction in either acquisition or clearance (Figure 3A-C, eTable 2). The interaction parameter in clearance $\log(h)$ was biased toward zero (no interaction), while the interaction parameter in acquisition $\log(k)$ was either over- or under-estimated. Overall, estimates under BALB were less biased than under the naive method. In particular, BALB was better in estimating the ratio of the two interaction parameters $\log(k/h)$ as the sampling interval became longer; in Figure 3, the estimates of BALB (green) stayed closer to the dashed slopes containing the true parameter values than the estimates of the naive method (red). When increasing the number of types, the biases under both methods remained at similar levels (Figure 3D-F, eTable 3). However, the CIs became narrower. In some instances, the CIs even excluded the true parameter values due to precise but biased estimation.

Regarding the estimation of the baseline rates, again, short sampling intervals (0.5 months) safeguarded against biased estimation; longer sampling intervals induced larger bias, while increasing the number of types did not influence the bias (eTables 4 and 5). The bias due to increasing length of sampling interval was comparable across the two methods. Of note, longer sampling intervals led to systematic underestimation of both baseline rates by both methods. Such underestimation is caused by the minimum-transition assumption, used in both the BALB and naive methods. When a complete carriage episode is missed, the true trajectory cannot be captured by a minimum-transition trajectory, as it would require acquisition and clearance of the same type. Both the numbers of acquisition and clearance events are then underestimated, which in turn leads to an underestimation of the baseline rates.

4.5.2 | Performance of BALB under heterogeneity in exposure

Next, we explored the ability of BALB to adjust for unobserved heterogeneity in exposure. In the setting with competition in acquisition but no interaction in clearance ($k = 0.5$, $h = 1$), BALB became increasingly biased in estimating the interaction parameter of acquisition with increasing heterogeneity when random effects were not included in the analysis (Figure 4A, eTable 6). The bias was toward more synergy; when unobserved heterogeneity was sufficiently large ($\alpha = 1$), competition in acquisition was even erroneously indicated as synergistic. Including random effects, BALB was able to correct for the unobserved heterogeneity. However, in the setting resembling homogeneity, the estimates of interaction in acquisition were slightly more biased toward stronger competition if random effects were allowed, likely due to overfitting. It is also noteworthy that the estimates of the interaction parameter in clearance did not seem to be affected by the amount of heterogeneity, which acts on the acquisition rates (Figure 4B, eTable 6). The other setting with competition in clearance but no interaction in acquisition ($k = 1$, $h = 2$) gave the same qualitative results (Figure 4C,D, eTable 6).

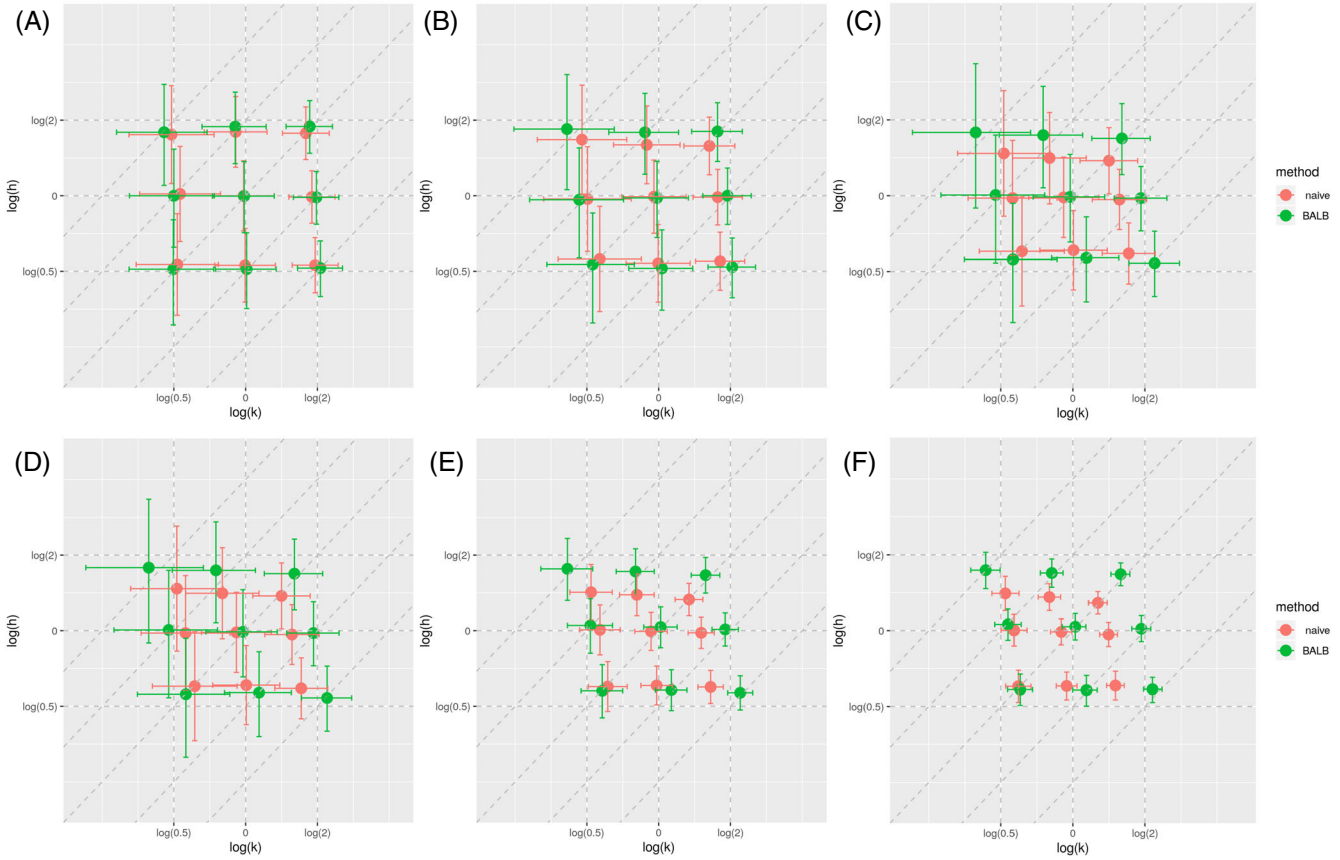

FIGURE 3 Estimates of BALB and the naive method in absence of heterogeneity in exposure. Estimates of the log-transformed interaction parameters, $\log(k)$ and $\log(h)$, obtained by BALB (green) and the naive method (red), in settings in absence of heterogeneity in exposure. The top row shows the settings with 2 types and increasing length of sampling intervals: (A) 0.5 months, (B) 1 month, and (C) 2 months. The bottom row shows the settings with 2-month sampling intervals and increasing number of types: (D) 2 types, (E) 4 types, and (F) 7 types. The vertical and horizontal dashed lines show the true values of $\log(k)$ and $\log(h)$, respectively. The dashed slopes through the intersections show the isoclines of $\log(k/h)$

5 | APPLICATION TO LONGITUDINAL PNEUMOCOCCAL CARRIAGE DATA

5.1 | Data collection

To further assess the performance of BALB, we analyzed a data set of pneumococcal carriage derived from 45 new-born infants in the Netherlands. Each infant was followed from birth up to at most 18 months of age. Nasopharyngeal samples were collected according to a decelerating sampling scheme in which samples were taken with 1-month intervals up to 8 months of age and subsequently at 10, 12, 15, and 18 months of age. Occasionally, additional samples were obtained if the infant experienced a respiratory infection. Nine infants were vaccinated with the 7-valent and 36 infants with the 10-valent pnemococcal conjugate vaccine.

Carriage of *S pneumoniae* and of individual pneumococcal serotypes was determined using both conventional diagnostic culture and molecular methods. The presence of 21 pneumococcal serotypes was determined as follows. For 14 serotypes (3, 6C, 9N, 10A, 11A, 12F, 15A, 15BC, 16F, 19A, 19F, 22F, 23A, and 33F), presence was identified both with molecular-based (qPCR) methods and culture. Carriage of the remaining 7 serotypes (17F, 21, 23B, 25F, 31, 35B, and 35F) was identified by culture only because no molecular-based method was available at the time of data collection. Positivity is here defined as positive by either one of the two methods. Samples positive for *S pneumoniae* for which the serotype could not be determined were assigned as non-typable (NT). Details of the detection methods are given elsewhere.[19]

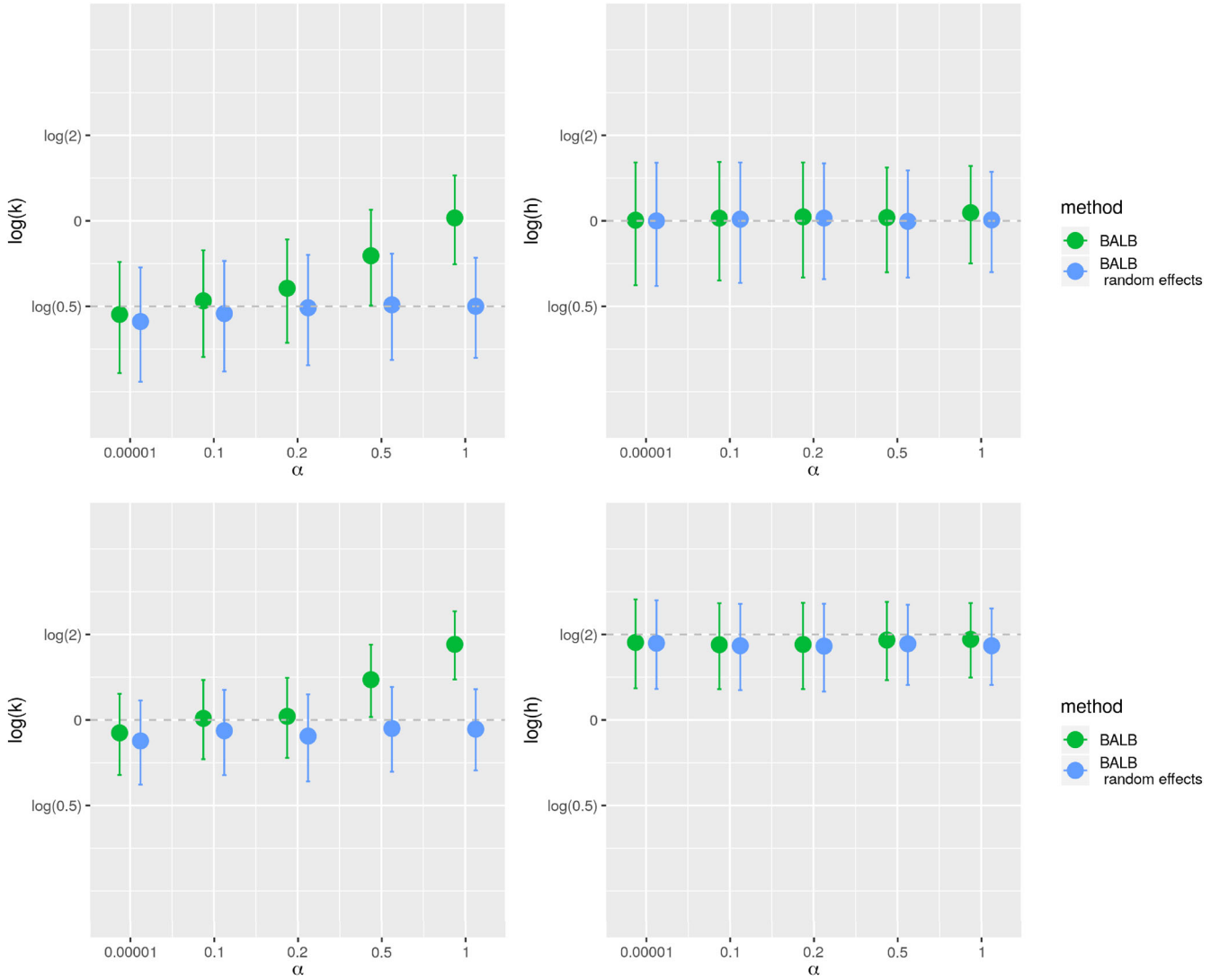

FIGURE 4 Estimates of BALB under increasing heterogeneity in exposure. Estimates of the log-transformed interaction parameters $\log(k)$ (A, C) and $\log(h)$ (B, D) obtained by BALB with (blue) and without (green) random effects in settings with 2 types, 1-month sampling interval and increasing amount of unobserved heterogeneity in exposure. The top row (A, B) shows the settings with competition in acquisition and no competition in clearance ($k = 0.5, h = 1$). The bottom row (C, D) shows the settings with no competition in acquisition and competition in clearance ($k = 1, h = 2$). The horizontal dashed lines show the true values of $\log(k)$ (left) and $\log(h)$ (right)

The data set comprises 538 samples, of which 109 (20% of 538) were additional samples obtained due to respiratory infections. Co-carriage of up to 4 serotypes was determined. In total, 335 (62% of 538) of the samples were positive for at least one serotype, among which 107 (32% of 335) were positive for multiple serotypes (see eFigure 1 in Supplementary Appendix F for the distribution of the number of types in co-carriage). Pneumococcal carriage was heterogeneous across the serotypes, ranging from 2.6% to 12.3% of the samples (eFigure 2 in Supplementary Appendix F). The proportion of samples positive for *S pneumoniae* also varied across individuals; which was 43.5% among the 50% of the study participants with the least carriage and 79.5% among the 50% of the study participants with the most carriage. This suggests the need to adjust for heterogeneity in exposure (or likewise, in predisposition to pneumococcal colonization) across individuals by including random effects.

5.2 | Methods

We analyzed the pneumococcal data set with the naive method and with the BALB method, with and without random effects. To ensure stable estimates, we assumed common baseline acquisition and clearance rates for the seven rarest

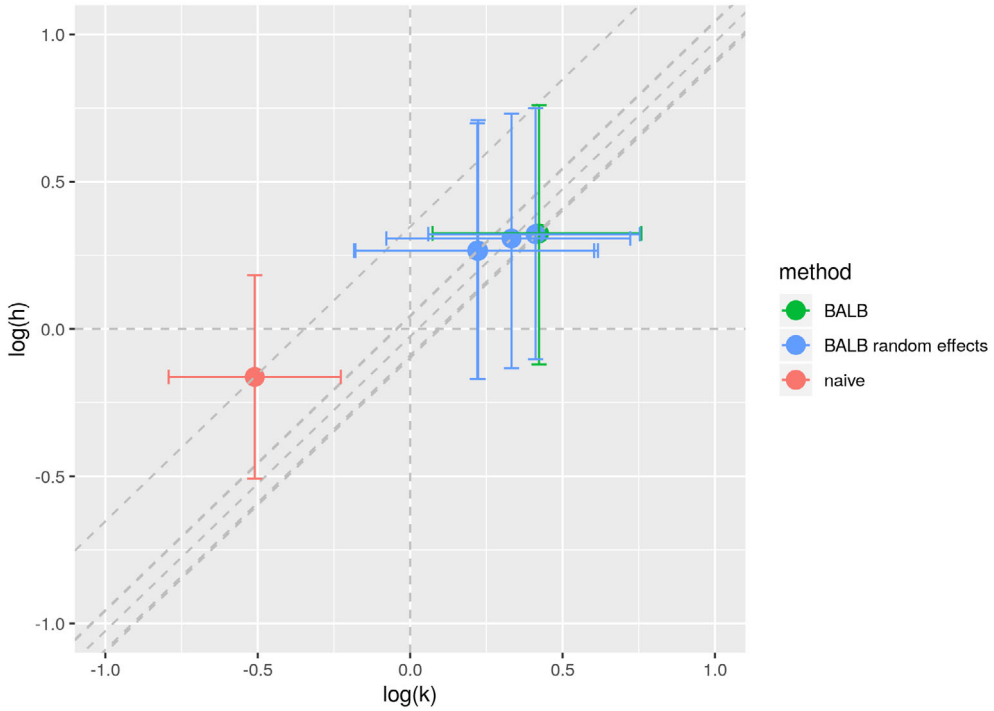

FIGURE 5 Estimates of BALB and the naive method for the pneumococcal application. Estimates of the log-transformed interaction parameters $\log(k)$ and $\log(h)$ obtained by the naive method (red), BALB with (blue) and without (green) random effects. Using wider priors for the variance α of the random effects by increasing the standard deviation of its half-normal prior from 0.01, 0.1, 1 to 10 resulted in smaller estimates of the interaction parameter of acquisition (blue circles more to the left)

serotypes (12F, 17F, 22F, 25F, 3, 31, and 9N) in all methods. For BALB, the same priors for the model parameters were used as in the simulation study, except for α, for which a half-normal distribution was assumed. For the half-normal prior distribution, we also explored the effect of increasing standard deviation (SD $= 0.01, 0.1, 1, 10$) while fixing the mean to zero.

5.3 | Results

The baseline rates estimated by the different methods were comparable, whereas the estimates of the interaction parameters were quite divergent (eTable 8, Figure 5). While the naive method indicated competition in acquisition and synergy in clearance, BALB with or without random effects suggested competition in clearance and synergy in acquisition. Similar to the simulation study, using wider priors for α in BALB with random effects led to smaller estimates of the interaction parameter of acquisition, corresponding to less synergy (or more competition). Simultaneously, the estimates of the interaction parameters of clearance became slightly smaller, corresponding to less competition.

The estimates of the log-transformed ratio of the two interaction parameters $\log(k/h)$ were more consistent across the different methods. The naive method yielded a large negative log-transformed ratio, indicating strong competition. The estimates of BALB were closer to zero. BALB without random effects found a small but positive log-transformed ratio. Allowing more heterogeneity across individuals turned the estimated log-transformed ratio to a negative value, indicating competition.

6 | DISCUSSION

In this article, we developed a new estimation method for inferring interactions between multiple types of the same pathogen from longitudinal data. The new method was developed to accommodate inference of between-serotype interactions from pneumococcal carriage data obtained using molecular-based detection methods, typically yielding increased

levels of co-carriage as compared to traditional culture-based methods. Multi-state models have difficulty in dealing with co-carriage states of more than two types due to the increased computational burden of exploring all possible acquisition and clearance events between discrete observation times. The new estimation method tackles the computational task by approximating the likelihood function of the parameters in a multi-state model with the likelihood of trajectories with a minimum number of transitions. To facilitate estimation, the approximate likelihood function was embedded in a Bayesian framework and implemented according to a computationally efficient Hamiltonian Monte Carlo scheme. The performance of the resulting BALB estimation method was demonstrated in a simulation study. Moreover, by applying BALB to a pneumococcal carriage data set, we were able to shed new light on the extent and mechanism by which pneumococcal serotypes compete during colonization.

Because the computational burden of inferring between-type interactions could already be high for models with only co-carriage states up to two types, many previous methods have taken recourse to some form of approximation to mitigate this burden. For example, methods based on data augmentation or maximum likelihood estimation have also relied on the minimum-transition assumption.[4,6] There are some approaches that relax this assumption but all previously considered multi-state models omit co-carriage states of more than two serotypes.[2,3,5,7,20] By capping the number of co-occurring serotypes to two, it is implicitly assumed that an individual already colonized with two types may not acquire any additional types before some are cleared first, which may induce a bias toward stronger competition. For the estimation method here, we chose to keep the minimum-transition assumption in order to be able to relax the limitation on the number of co-occurring serotypes. This enables estimation of between-type interaction from molecular-based pneumococcal carriage data. The loss in accuracy resulting from the minimum-transition assumption was found to be justified in our specific application, given the comparatively frequent sampling in relation to acquisition and clearance of pneumococcal serotypes.

We evaluated the performance of BALB using simulated data. As a benchmark, we used a naive estimation method which simply imputes transition times midway between the observation times. Although the simple method is computationally undemanding, it may be more prone to bias as midpoint transitions may systematically shorten or lengthen episodes of carriage or non-carriage relative to the true colonization history.

The simulation study also revealed under which circumstances BALB may perform suboptimally. Long sampling intervals led to biases in the estimates of the model parameters; the baseline rates were underestimated, interaction in clearance was biased toward no competition, and both competition and synergy in acquisition could be either over- or underestimated. Nevertheless, the ratio of the two interaction parameters remained relatively unbiased even with long sampling intervals. The underestimation of the baseline rates is due to missing of a portion of complete carriage episodes between consecutive observation times, which cannot be accounted for by the minimum-transition assumption. It is less clear what determines the direction of biases in the interaction parameters. Our hypothesis is that the rates of clearance are more easily biased than the rates of acquisition, and likewise, the estimation of interaction in clearance is more difficult. While the average duration of a carriage episode is around the same order of magnitude as the sampling interval, periods of non-carriage are much longer, facilitating precise estimation of acquisition rates. As the ratio of the two interaction parameters is more easily identifiable, it is conceivable that the method tends to compensate for the bias in the interaction parameter of clearance with the interaction parameter of acquisition. More investigation is needed to substantiate this hypothesis. While it would be ideal to estimate all parameters without bias, it could be sufficient to accurately estimate the ratio of the interaction parameters in situations where the total amount of competition is more relevant. As we have shown previously, the ratio of the interaction parameters in acquisition and clearance constitutes a natural summary measure of the interactions in these two modes and contains predictive value for serotype replacement.[7,21]

The simulation study also showed that BALB was capable of adjusting for unobserved heterogeneity in exposure or predisposition to pneumococcal colonization by including individual-specific random effects. In the context of the motivating data on pneumococcal carriage in infants, exposure of pneumococcal carriage may be heterogeneous due to differences in, for example, the number of siblings and attendance to daycare. When unadjusted for, unobserved factors that increase the risk of carriage irrespective of type could induce spurious positive associations between different serotypes, masking possible competition between serotypes.[21] In the simulation study, adjustment for unobserved heterogeneity by including random effects worked satisfactorily. However, it could be more challenging in reality. Possible challenges include correct specification of the distribution of random effects and the mechanism through which unobserved heterogeneity acts. Therefore, measuring and modeling possible confounders remain pivotal.

After validation on simulated data, the proposed estimation method was used to study interactions between pneumococcal serotypes based on a data set of pneumococcal carriage obtained using molecular-based detection methods. In line with previous results, which were all derived from culture-based data,[1,2,4,6,7] we found evidence for competition between

serotypes. Since the detection method that we used to determine the composition of co-carrying serotypes in samples is more sensitive than the culture-based methods used in previous studies, we anticipated between-type competition to be weaker than previously estimated.[13] Nevertheless, the estimated effect size was surprisingly low. An additional explanation is that the serotypes with strong competitive ability were already removed from the vaccinated study population, as the vaccine targets serotypes that dominated in carriage, such as 6B, 14, 19F, and 23F, which are arguably also the strong competitors.[9,22,23] It was also surprising that competition was found in clearance and not in acquisition in this data set. Although previous carriage studies have identified competition in acquisition to be the main mechanism, some of them as well as experiments in mouse models have hinted on competition in clearance.[4,7,8] More studies on other data sets are needed to validate these results.

The analysis of between-type interactions on this data set has some shortcomings. Most importantly, the sample size was small (45 individuals). To put this number into perspective, most other longitudinal carriage data sets used for estimating pneumococcal interactions have consisted of more than a hundred subjects.[2-7,20] In addition, at the time of data collection, only 14 of the 21 serotypes could be identified by molecular-based methods. The seven remaining serotypes were thus detected using culture-based methods only. The level of unobserved heterogeneity may also have complicated identification of between-type interactions. Inclusion of random effects may have only been able to partly account for this. Adjustment based on the collected background information of the study participants and seasonal effects could have further strengthened the analysis.

In the future, the proposed method could be applied to other data sets to further elucidate interaction between pneumococcal serotypes. In principle, the approach remains valid as long as the transition rates are constant between observation times. For instance, it is possible to model type-specific interaction parameters, which may enable detection of possible differences in competitive ability across pneumococcal serotypes, a topic that has only been touched upon in previous studies.[2,6,7] When factors such as age, season, and medication use are relevant to pneumococcal carriage,[6,20] they could be included as covariates. While introducing more parameters may make the model more realistic, it should be done with caution, as identification of the large number of parameters is challenging when the amount of data is limited.

As methods to detect and characterize multi-strain pathogens will continue to improve and become more accessible, we expect more opportunities to study the interactions between pneumococcal serotypes or strains of other pathogens, such as the human papillomavirus and *P falciparum*.[24,25] Existing statistical and computational methods may need to be further refined to accommodate analysis of the resulting higher-resolution data. Whenever more accurate measurements of pathogen types or strains are made at regular intervals that are sufficiently short to guarantee that repeated acquisition and clearance in between measurements is unlikely, the inferential approaches based on minimum-transition trajectories like the one presented here can be used to study pathogen interactions. In turn, new knowledge of between-strain interactions may help to better understand and improve the impact of key preventive and therapeutic interventions (eg, vaccination and antibiotics) against multi-strain pathogens.[10,26]

ACKNOWLEDGEMENTS

This work was supported by grant S/113005/01/PT (Prometheus project) through the Strategic Programme from the National Institute for Public Health and the Environment (RIVM) of the Netherlands.

DATA AVAILABILITY STATEMENT

All relevant data are within the article and its supplementary material.

ORCID

Irene Man https://orcid.org/0000-0003-3177-6904

REFERENCES

1. Lipsitch M, Dykes J, Johnson S, et al. Competition among *Streptococcus pneumoniae* for intranasal colonization in a mouse model. *Vaccine*. 2000;18(25):2895-2901.
2. Melegaro A, Choi Y, Pebody R, Gay N. Pneumococcal carriage in United Kingdom families: estimating serotype-specific transmission parameters from longitudinal data. *Am J Epidemiol*. 2007;166(2):228-235.
3. Hoti F, Erästö P, Leino T, Auranen K. Outbreaks of *Streptococcus pneumoniae* carriage in day care cohorts in Finland—implications for elimination of transmission. *BMC Infect Dis*. 2009;9(1):102.
4. Auranen K, Mehtälä J, Tanskanen A, and Kaltoft MS. Between-strain competition in acquisition and clearance of pneumococcal carriage—epidemiologic evidence from a longitudinal study of day-care children. *Am J Epidemiol*. 2010;171(2):169-176.

5. Erästö P, Hoti F, Auranen K. Modeling transmission of multitype infectious agents: application to carriage of *Streptococcus pneumoniae*. *Stat Med*. 2012;31(14):1450-1463.
6. Lipsitch M, Abdullahi O, D'Amour A, 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(4):510.
7. Mehtälä J, Antonio M, Kaltoft MS, O'Brien KL, Auranen K. Competition between *Streptococcus pneumoniae* strains: implications for vaccine-induced replacement in colonization and disease. *Epidemiology*. 2013;24(4):522-529.
8. Trzciński K, Li Y, Weinberger DM, et al. Effect of serotype on pneumococcal competition in a mouse colonization model. *mBio*. 2015;6(5):e00902-e00915.
9. Weinberger DM, Malley R, Lipsitch M. Serotype replacement in disease after pneumococcal vaccination. *Lancet*. 2011;378(9807):1962-1973.
10. Hausdorff WP, Hanage WP. Interim results of an ecological experiment—conjugate vaccination against the pneumococcus and serotype replacement. *Hum Vaccin Immunother*. 2016;12(2):358-374.
11. Lewnard JA, Hanage WP. Making sense of differences in pneumococcal serotype replacement. *Lancet Infect Dis*. 2019;19(6):e213-e220.
12. Turner P, Hinds J, Turner C, et al. Improved detection of nasopharyngeal cocolonization by multiple pneumococcal serotypes by use of latex agglutination or molecular serotyping by microarray. *J Clin Microbiol*. 2011;49(5):1784-1789.
13. Satzke C, Dunne EM, Porter BD, Klugman KP, Mulholland EK, Project Group P. The PneuCarriage project: a multi-centre comparative study to identify the best serotyping methods for examining pneumococcal carriage in vaccine evaluation studies. *PLoS Med*. 2015;12(11):e1001903.
14. Olwagen CP, Adrian PV, Madhi SA. Comparison of traditional culture and molecular qPCR for detection of simultaneous carriage of multiple pneumococcal serotypes in African children. *Sci Rep*. 2017;7(1):1-9.
15. Weinberger DM, Dagan R, Givon-Lavi N, Regev-Yochay G, Malley R, Lipsitch M. Epidemiologic evidence for serotype-specific acquired immunity to pneumococcal carriage. *J Infect Dis*. 2008;197(11):1511-1518.
16. Granat SM, Ollgren J, Herva E, Mia Z, Auranen K, Mäkelä PH. Epidemiological evidence for serotype-independent acquired immunity to pneumococcal carriage. *J Infect Dis*. 2009;200(1):99-106.
17. Moler C, Van Loan C. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Rev*. 1978;20(4):801-836.
18. Andersen PK, Keiding N. Multi-state models for event history analysis. *Stat Methods Med Res*. 2002;11(2):91-115.
19. Wyllie A, Bogaert D, Prevaes S, et al. Carriage of *Streptococcus pneumoniae* in the first 18 months of life. *Manuscript in preparation*.
20. Numminen E, Chewapreecha C, Turner C, et al. Climate induces seasonality in pneumococcal transmission. *Sci Rep*. 2015;5:11344.
21. Man I, Wallinga J, Bogaards JA. Inferring pathogen type interactions using cross-sectional prevalence data: opportunities and pitfalls for predicting type replacement. *Epidemiology*. 2018;29(5):666-674.
22. Weinberger DM, Trzciński K, Lu YJ, et al. Pneumococcal capsular polysaccharide structure predicts serotype prevalence. *PLoS Pathog*. 2009;5(6):e1000476.
23. Vissers M, Wijmenga-Monsuur AJ, Knol MJ, et al. Increased carriage of non-vaccine serotypes with low invasive disease potential four years after switching to the 10-valent pneumococcal conjugate vaccine in The Netherlands. *PLoS One*. 2018;13(3):e0194823.
24. Ranjeva SL, Baskerville EB, Dukic V, et al. Recurring infection with ecologically distinct HPV types can explain high prevalence and diversity. *Proc Natl Acad Sci*. 2017;114(51):13573-13578.
25. Lerch A, Koepfli C, Hofmann NE, et al. Longitudinal tracking and quantification of individual *Plasmodium falciparum* clones in complex infections. *Sci Rep*. 2019;9(1):1-8.
26. Colijn C, Corander J, Croucher NJ. Designing ecologically optimized pneumococcal vaccines using population genomics. *Nat Microbiol*. 2020;5(3):473-485.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Man I, Bogaards JA, Makwana K, Trzciński K, Auranen K. Approximate likelihood-based estimation method of multiple-type pathogen interactions: An application to longitudinal pneumococcal carriage data. *Statistics in Medicine*. 2022;41(6):981-993. doi: 10.1002/sim.9305