Biometrics 69, 748–757
September 2013

DOI: 10.1111/biom.12040

Estimating the Transmission Dynamics of *Streptococcus pneumoniae* from Strain Prevalence Data

Elina Numminen,[1,*] Lu Cheng,[1] Mats Gyllenberg,[1] Jukka Corander[1,2,3]

[1]Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68, 00014 Helsinki, Finland
[2]Helsinki Institute for Information Technology, University of Helsinki, P.O. Box 68, 00014 Helsinki, Finland
[3]Department of Mathematics, Åbo Akademi University 20500 Åbo, Finland
**email:* elina.numminen@helsinki.fi

Summary. *Streptococcus pneumoniae* is a typical commensal bacterium causing severe diseases. Its prevalence is high among young children attending day care units, due to lower levels of acquired immunity and a high rate of infectious contacts between the attendees. Understanding the population dynamics of different strains of *S.pneumoniae* is necessary, for example, for making successful predictions of changes in the composition of the strain community under intervention policies. Here we analyze data on the strains of *S. pneumoniae* carried in attendees of day care units in the metropolitan area of Oslo, Norway. We introduce a variant of approximate Bayesian computation methods, which is suitable for estimating the parameters governing the transmission dynamics in a setting where small local populations of hosts are subject to epidemics of different pathogenic strains due to infections independently acquired from the community. We find evidence for strong between-strain competition, as the acquisition of other strains in the already colonized hosts is estimated to have a relative rate of 0.09 (95% credibility interval [0.06, 0.14]). We also predict the frequency and size distributions for epidemics within the day care unit, as well as other epidemiologically relevant features. The assumption of ecological neutrality between the strains is observed to be compatible with the data. Model validation checks and the consistency of our results with previous research support the validity of our conclusions.

Key words: Bayesian inference; Continuous-time Markov chains; Epidemiology; Multi-strain models; Transmission dynamics.

1. Introduction

The bacterium *Streptoccus pneumoniae* is a common colonizer of the upper respiratory tract, the *nasopharynx*, of humans. While in most cases asymptomatic and harmless, the colonization might progress to invasive diseases, such as sepsis, meningitis and pneumonia. Being both the source of the horizontal spread of the pathogen and the predecessor of the diseases, the process of asymptomatic colonization plays a key role in determining the epidemiology of pneumococcal diseases. So far it has been well established that young children belong to the risk groups for the pneumococcal diseases, and also that colonization is more prevalent among them (Bogaert, De Groot, and Hermans, 2004). Especially high levels of colonization have been observed in children that attend day care centers (DCCs) (Dunais et al., 2003), because of enhanced transmission between DCC attendees. Furthermore, Huang, Finkelstein, and Lipsitch (2009) suggested that the level of DCC attendance actually influences the overall community-level prevalence of pneumococci.

Conjugate vaccines have been recommended for protection against pneumococcal infections to those at greatest risk. These vaccines are highly effective against both the colonization and disease of a certain subset, typically seven, of the pneumococcal strains. However, pneumococcae are highly diverse bacteria, for which over 90 serotypic strains are known. Studies monitoring the population-level effects of vaccines (Kellner et al., 2008; Vestrheim et al., 2010) clearly report replacement of vaccine-strains by non-vaccine-strains over few years, so that eventually only the strain composition is altered, but no change is achieved in the prevalence of the carriage of pneumococcal strains. As the non-vaccine strains benefit from the decline of the vaccine-strains, the observations indicate the presence of between-strain competition in pneumococcal communities. This motivates study of the ecological interactions and the transmission of pneumococcal strains, as an understanding of these processes would help to explain the observations and allow for a precise evaluation of the utilities and implications of vaccination policies.

In this article, we address the question of between-strain competition in pneumococcal populations while studying the transmission of colonization within children that attend DCCs. Among small groups of hosts, such as DCCs, the processes of transmission and colonization are highly stochastic. Despite the randomness, localized epidemics can have important population-level implications. For instance, Ball, Mollison, and Scalia-Tomba (1997) argued that localized transmission provides an amplifying effect for the global prevalence of an epidemic and Hagenaars, Donnelly, and Ferguson (2004) demonstrated that the asynchrony of small localized epidemics facilitates the global persistence of a pathogen in the population.

Mathematical models for stochastic transmission within small social groups are typically referred to as *household models*. Statistical inference on epidemics for such models has been conducted, for instance, using longitudinal data (Hoti et al., 2009) or final-size distributions of infected cases (Brooks-Pollock et al., 2011). While methods analyzing longitudinal data typically assess the plausibility of different model parameters by conditioning on the observed event histories, methods using final-size distributions evaluate different parameters by comparing the total numbers of infections within households predicted by the model and observed in the data.

In this article, we analyze cross-sectional data on pneumococcal strains carried in a total of 611 children from 29 DCCs. In particular, we perform Bayesian inference for a transmission model, under which new epidemics are continuously initiated in the DCC from the community, but since the strains compete for colonizing hosts, the outcomes of the epidemics are suppressed. The presented analysis is related to the final-size approaches, since the likelihood is approximated by conditioning on the characteristics of recurrent epidemics of *different strains* within the DCCs, reflected by the states of DCCs in the data. To perform the analysis, we used the methods of approximate Bayesian computation (Beaumont, 2010), replacing the evaluation of the likelihood function with direct simulations from the generating model.

Regarding the results, we obtained unimodal posterior distributions for the relative rates governing the transmission dynamics. These results allowed us to assess the strength of the competition between the strains, and also to predict the unobserved properties of the system, and behavior of the system in time, once the time scale was adjusted through the application of external knowledge. Based on our model validation checks and comparisons with estimates obtained in another study, our method appears to be successful in capturing the relevant features of the transmission dynamics. Thus it is demonstrated that strain-diversity distributions can be very informative of the transmission processes, once ecological interactions between the strains are included in the model.

We are not aware of previous studies of comparable epidemiological inference utilizing cross-sectional observations on strain diversity. Data on bacterial genotypes has previously been used for epidemiological inference (Luciani et al., 2009). However, as genotypes evolve much faster than serotypes, analysis of such data must account for both the ecological and evolutionary processes, while a model for strain data can, under the appropriate circumstances, neglect the evolutionary processes acting on the studied organism.

2. Data and Model

2.1. *Pneumococcal Carriage Data*

The data we analyze in this article originate from a point-prevalence study on pneumococcal strains carried in children attending DCCs in the capital region of Oslo, Norway. This article was performed in 2006, before the introduction of conjugate vaccines in the Norwegian national vaccination program. A detailed description of the data is found in Vestrheim et al. (2008), and results from a follow-up study are given in Vestrheim et al. (2010).

In the article, 611 healthy children from 29 DCCs were investigated regarding their pneumococcal carriage. A nasopharyngeal swab was taken from each participating child to identify the strains carried. In total, 33 different strains were observed among these swabs. Labeling both the DCCs and carried strains with integers, the data can be represented in the form $\{\boldsymbol{DCC}_i, \boldsymbol{C}_i\}, i = 1, \ldots, 611$, where $\boldsymbol{DCC}_i$ defines the label of the DCC of the i'th individual and $\boldsymbol{C}_i$ is a binary vector for which $\boldsymbol{C}_i(j) = 1$, $j = 1, \ldots, 33$, if individual i is a carrier of strain j, and $\boldsymbol{C}_i(j) = 0$ otherwise. On average, 21 children from each DCC were studied, the minimum and maximum numbers of sampled children from a single DCC being 6 and 37, respectively. Unfortunately, the total number of children attending each DCC was not available for our modeling purposes. Nevertheless, the average number of children attending these DCCs was reported to be 53.1 by Vestrheim et al. (2008).

The diversity distribution of the serotypic strains over all the collected swabs, shown in Figure 1, manifests a very diverse population. The majority of strains are relatively rare, and only 12% of the total number of colonized individuals are colonized by the most common strain. A high diversity of strains was also observed within each DCC. On average, 8.1 different strains were observed among individuals from the same DCC, and on average the most frequently observed strain in each DCC was colonizing 31% of them. In Web Figure 1, we illustrate the within-DCC diversities of strains in more detail. Regarding the prevalence of colonization, 77.7% of the studied children were colonized by pneumococci. Out

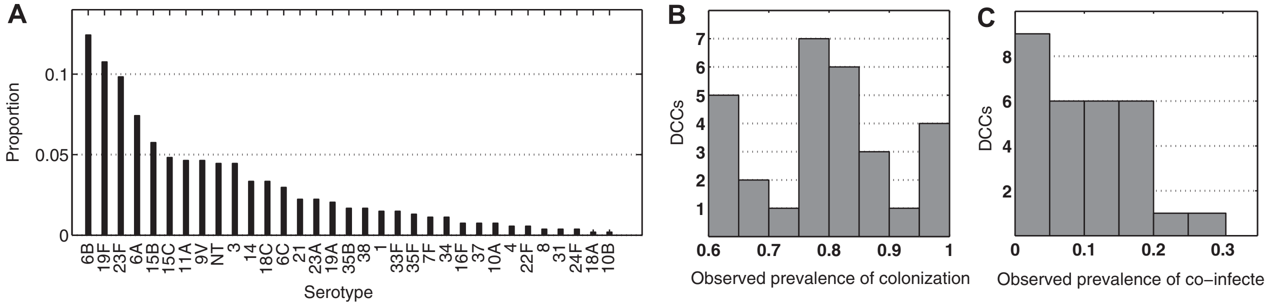

Figure 1. (A) The overall distribution of the different strains in the data as proportions of individuals that were colonized with the particular strain out of the total number of colonized individuals. In total 611 children were investigated. (B) and (C) The distributions over the 29 studied DCCs for the observed prevalence of colonized individuals and the observed prevalence of individuals colonized with multiple strains as proportions of the total number of sampled individuals from each DCC.

of the colonized children, 12.6% were colonized with multiple strains, that is, co-infected, usually by two strains. Among the groups of studied children from each separate DCC, the proportion of individuals that was colonized varied from 0.609 to 1, and the proportion colonized with several strains varied between 0 and 0.3. The between-DCC variation in these two observables is also illustrated in Figure 1. The prevalence of co-infection within a DCC was not particularly correlated with the number of different strains observed in the DCC, the correlation coefficient between the two being only 0.2368. The reason for this is likely to be that all of the DCC' were similarly diverse to a sufficient degree regarding co-infection. The DCC-wise observations are affected by the number of individuals investigated in each, as well as the unknown total numbers of attendees. In the following discussion, we assume that the DCCs were of an equal size that was also constant over time.

2.2. *Stochastic SIS-Model for the Spread of Strains within a DCC*

We now define the stochastic model for the transmission dynamics of N^s strains among a single day care centre consisting of N^{ind} individuals. The system state at each instance of time is an $N^{\text{ind}} \times N^s$ -matrix $\boldsymbol{I}(t)$, for which $\boldsymbol{I}_{ij}(t) = 1$, if individual i is colonized by strain j at time t, and $\boldsymbol{I}_{ij}(t) = 0$ otherwise. The individuals are assumed to be subject to a constant force of infection from a large community, and the prevalences of different strains to be stable within the community. The assumption of the community-level diversity distribution of strains to be stable agrees with previous empirical studies, for example, Hanage (2010). We parameterize the transmission process using four parameters: β, Λ, θ, and γ, where β and Λ account for the hazards of infection from the DCC and from the community, θ scales for the probability of co-infection and γ is the rate of clearance of an infection.

The core assumptions leading to such parameterization are explained below. Firstly, the DCC attendees are assumed to have encounters with other hosts that can lead to transmission of colonization. Such encounters include contacts with other members of the DCC and with individuals outside the DCC, but the rate of encountering other DCC members, denoted with c_1, is different from the rate at which DCC members have contacts with other individuals in the community outside the DCC, which we denote by c_2. While the proportion P of individuals colonized by the pneumococci in the whole community is assumed to be constant, we assume that the number of colonized individuals within a DCC, denoted by $I^{tot}(t)$, is a stochastic process with transition probabilities that depend on the current colonization status of each individual in the DCC. Assuming further that, given a contact between a susceptible individual and an infected host, the probability of transmission is p, then the probability per time unit to acquire an infection from the community is $c_2 pP$ (i.e., a constant). The corresponding probability for a susceptible DCC attendee to acquire a colonization by some strain from another DCC attendee depends on $\mathbf{I}(t)$, and this equals $c_1 p((I^{tot}(t))/(N^{\text{ind}}))$. Since the actual transmission processes depend on the products of the parameters, we parameterize the model using $\beta := c_1 p$ and $\Lambda := c_2 pP$. Notice that the two parameters have different interpretations, since P was included in Λ.

When modeling the transmission process of several different strains, similar assumptions hold: we denote the proportion of colonized individuals in the community who are carrying the strain s with P^s, and we assume that P^s is constant in time for all s. Now, we assume that if a susceptible individual host has an encounter with an individual in the community, the probability that strain s is transmitted is pPP^s. To define the corresponding probability that strain s is transmitted, given an encounter with another individual from the DCC, we use the following notation:

$$E_s(\boldsymbol{I}(t)) = \sum_{\{i \,:\, \boldsymbol{I}_{is}(t)=1\}} \left(\frac{\boldsymbol{I}_{is}(t)}{N^{\text{ind}} - 1} \frac{1}{\sum_{j=1}^{N^s} \boldsymbol{I}_{ij}(t)} \right). \tag{1}$$

We then assume that the probability that strain s is transmitted, given that a susceptible DCC attendee has a within-DCC contact at time t, is $pE_s(\boldsymbol{I}(t))$. Notice that $E_s(\boldsymbol{I}(t))$ is the probability of sampling strain s in a hierarchical random experiment in which first another DCC attendee is sampled, and if this individual is colonized, a random strain is sampled from the strains colonizing that individual. The $N^{\text{ind}} - 1$ term in the denominator accounts for the fact that the randomly sampled DCC attendee can not be the susceptible individual herself.

We assume that hosts can be colonized by several strains simultaneously, but to account for competition between the strains, we use a further parameter θ to scale the probability of becoming colonized with a further strain, once already colonized with one or many strains of *another type*. We also assume only one colonizing strain is transmitted in a transmission event. Finally, we assume that the duration of carriage of a strain is exponentially distributed, with a mean $1/\gamma$.

Given the assumptions and notations above, the model is defined as a continuous-time Markov process, which has the following transition probabilities for a small time increment δt:

$$\begin{aligned}
&Pr\left(\boldsymbol{I}_{is}(t+\delta t) = 1 | \boldsymbol{I}_{is}(t) = 0\right) = \beta E_s(\boldsymbol{I}(t)) + \Lambda P^s + o(\delta t), \\
&\qquad\qquad \text{if} \quad \sum_{j=1}^{N^s} \boldsymbol{I}_{ij}(t) = 0, \\
&Pr\left(\boldsymbol{I}_{is}(t+\delta t) = 1 | \boldsymbol{I}_{is}(t) = 0\right) = \theta\left(\beta E_s(\boldsymbol{I}(t)) + \Lambda P^s\right) + o(\delta t)\right), \\
&\qquad\qquad \text{if} \quad \sum_{j=1}^{N^s} \boldsymbol{I}_{ij}(t) > 0 \quad \text{and} \quad \boldsymbol{I}_{is} = 0, \\
&Pr\left(\boldsymbol{I}_{is}(t+\delta t) = 0 | \boldsymbol{I}_{is}(t) = 1\right) = \gamma + o(\delta t).
\end{aligned} \tag{2}$$

The model is a stochastic variant of a standard SIS-model of pathogen spread (Anderson and May, 1991) with multiple co-circulating strains and a possibility of co-infection. The dynamics of the model are determined by defining the model parameters $(\beta, \Lambda, \theta, \gamma)$, N^{ind}, N^s and P^s for every $s = 1, \ldots, N^s$. In the parameter estimation process that follows, we set

$N^{\text{ind}} = 53$, which was the mean number of attendees in the studied DCCs and $N^s = 33$, which was the total number of different strains observed in the data. For $\boldsymbol{P}^s$ we use the overall serotype distribution in the data, shown in Figure 1.

Our transmission model contains a few further implicit assumptions. Firstly, through the construction used to define the probability in Equation (1) it is assumed that individuals colonized with one or many strains are equally infective. Also, as seen from Equation (2), the rate of clearance of a particular strain in a host is unaffected by the number of strains colonizing that host. This assumption is in agreement with the results in Auranen et al. (2010) on pneumococcal inter-strain competition, stating that the inhibition of colonization is the primary mechanism of competition, and that once two strains colonize the nasopharynx, the rate of clearance is not particularly accelerated. Finally, the model assumes ecological neutrality between the strains (Hubbel, 2001), in the sense that rates of events do not depend on the identities of the strains involved.

3. Inference

3.1. *Approximate Bayesian Computation Approach*

For the inference, we employ the tools of approximate Bayesian computation (ABC), discussed by, for example Beaumont (2010). In ABC, a sample from the posterior distribution is generated by approximating the likelihood function with simulations from the model: a candidate parameter $\boldsymbol{\varphi}$, sampled from the proposal distribution, is accepted to the posterior sample if the simulation results $\boldsymbol{D'}$ generated with $\boldsymbol{\varphi}$ are close enough to the observed data $\boldsymbol{D}$. The closeness can be assessed by defining a vector of summary statistics $S(\boldsymbol{D})$, distance measure ρ, and a threshold value for the distance, ε. The target distribution of ABC-inference is thereafter:

$$f(\varphi|\rho(S(\boldsymbol{D}), S(\boldsymbol{D'})) < \varepsilon). \tag{3}$$

For efficient implementation of ABC we use a variant of sequential ABC method (Sisson, Fan, and Tanaka, 2007) that has partial rejection control property. We also adjust the tolerance ε and the proposal distribution for parameters adaptively. For a detailed description of the computations, see the Web Supplementary Material.

3.2. *Simulation of the Model*

When using a cross-sectional dataset, parameters can be inferred only relative to each other. Therefore, when simulating, we set $\gamma = 1$, and estimate the other parameters relative to γ. We use the Gillespie method (Gillespie, 1976) for simulating random realizations from the transmission model. We start each simulation with no individuals being colonized and a constant force of infection from the community, defined by Λ. We perform 29 independent simulations for each proposed vector of the model parameters, each with $N^{\text{ind}} = 53$. Each simulation is run until time T, which is chosen such that $\boldsymbol{I}(T)$ should be sampled from the stationary distribution of $\boldsymbol{I}(t)$. Given the configurations of the 29 simulations at time T, we randomly assign the per-DCC sample sizes in the data among the 29 system states, and, given the sample sizes, sample the observed individuals from each simulated DCC without replacement. We denote the model parameters to be estimated by $\boldsymbol{\varphi} = (\beta, \Lambda, \theta)$ and the outcome of the described stochastic simulation procedure with $\boldsymbol{D'} \sim \Phi(\bullet \mid \boldsymbol{\varphi})$, where first 29 outcomes of the transmission processes with same parameters $\boldsymbol{\varphi}$ are generated and given the set of outcomes, the observation model is applied on that. This procedure ensures that $\boldsymbol{D'}$ consists of 611 observations on colonization configurations of individuals for each candidate set of parameters values, as in the data, and also that the numbers of observations are distributed between the DCCs in a similar way to their distribution in the data.

In the analysis, we used $T = 10$ as the simulation time, but performed exactly the same analysis using $T = 20$ to investigate the robustness of the results to the simulation time. As a further sanity check that the simulation time is adequate for different parameters to reach stationary distribution, we chose 10 different combinations of model parameters, and simulated the distributions of the four summary statistics listed in Section 3.3. with two different simulation times, 10 and 20. The distributions of the summary statistics were found to be highly congruent between the simulation times. An example of predicted Shannon index, with model parameters shown in Web table 1, is shown in Web Figure 2.

3.3. *Summary Statistics and the Distance Measure*

We approximate the likelihood function by assessing the probability of the distribution of DCC-specific summary statistics, like the ones shown in Figure 1 in panels B and C. In particular, we calculate the following summary statistics for each DCC, based on the strains carried by the individuals who were assigned the status of being observed:

1. Shannon index of diversity (Peet, 1974) of the distribution of observed strains.
2. Number of different strains.
3. Prevalence of carriage among the observed individuals.
4. Prevalence of multiple infections among the observed individuals.

Thus the dimensionality of the data $\boldsymbol{D}$ and the simulated datasets is reduced by projecting the observations into a 4×29 dimensional space, and the summary statistics considered are independent of the actual labels of the strains. To assess the discrepancy of the simulated $\boldsymbol{D'}$ from the data, we calculate for each summary $k = 1, \ldots, 4$, the L^1-distance between empirical distribution function of the summary, obtained from the simulation results, $F^k(x)$, and the empirical distribution function for the same summary, obtained from the real data $\hat{F}^k(x)$:

$$d_k = \int |F^k(x) - \hat{F}^k(x)| \, \mathrm{d}x. \tag{4}$$

The vector of corresponding discrepancies $\boldsymbol{d} = (d_1, d_2, d_3, d_4)$ is finally used as a proxy when approximating the likelihood function. By denoting the vector of tolerances with $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$, we require each d_k to be smaller than a particular tolerance value ε_k, for the simulation to yield an acceptance, that is, *to fall within the tolerance* $\boldsymbol{\varepsilon}$. Pritchard et al. (1999) also used a similar approach, requiring each summary statistic, after normalization, to be within a fixed

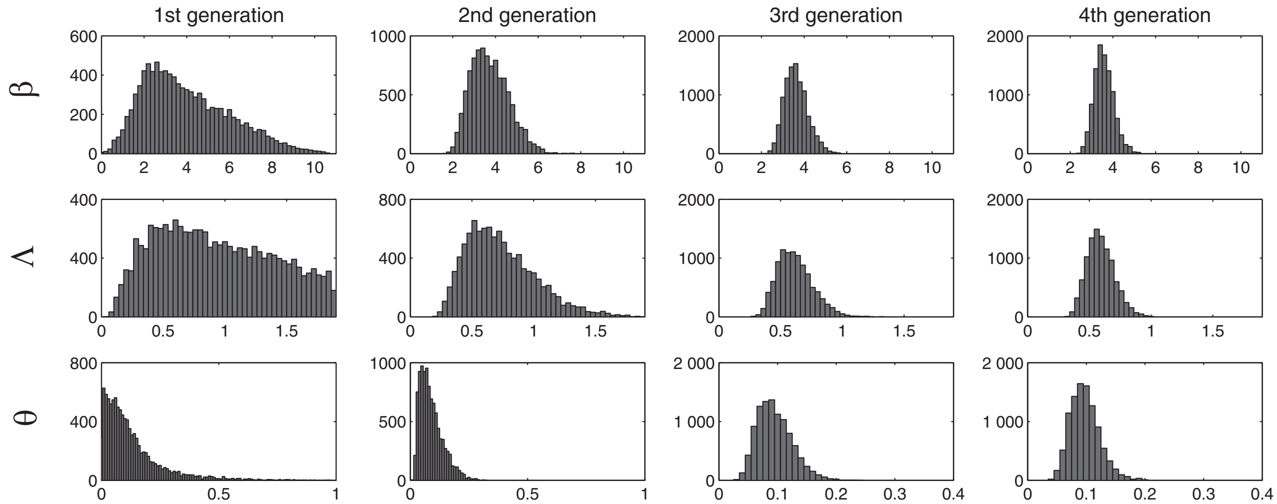

Figure 2. The columns in the figure show the approximated posterior distributions of the model parameters β, Λ, and θ in the consecutive sequential ABC-runs. The last column illustrates the distribution obtained in the fourth run, from which the summaries of the posterior were obtained, and according to which the posterior predictive simulations were performed.

$\boldsymbol{\varepsilon}$-distance from the observed statistic. Such approaches enhance the requirement that the simulated and the observed datasets are similar with respect to all of the desired features *simultaneously*. In principle, the likelihood of the data could also be defined by generalizing the approach in Stone et al. (2008) for multiple co-circulating strains with the possibility of co-colonization. However, evaluation of the likelihood would in practice be intractable (or extremely tedious), owing to the substantially higher dimension of the state-space and incomplete observations from each DCC.

3.4. *Prior Distributions*

Based on training simulations, we assigned the following prior distributions: $\beta \sim \text{Uniform}(0, 11)$, $\Lambda \sim \text{Uniform}(0, 2)$, and $\theta \sim \text{Uniform}(0, 1)$. We run 10,000 further simulations from these prior distributions to set the tolerance for the first sequential ABC-run, according to the criteria described in the Web Supplementary Materials.

4. Results

4.1. *Posterior Distributions of Parameters*

In Figure 2, we show the marginal posterior distributions of the estimated model parameters in four consecutive ABC-generations. The figure shows that the prior distributions were adequate in covering the support of the posterior. During the fourth sequential run, the time to reach the sample size increased almost threefold, and also a slight change in the marginal distributions was observed between the third and the fourth marginal distributions, so the sequential learning procedure was terminated after the fourth run. In Table 1, we summarize the posterior. Results from the analysis performed with double the simulation times are also shown. Comparison of the summaries of the two posteriors suggests that $T = 10$ was already a long enough simulation time to produce a sample from the stationary distribution of the process.

All of the reported estimates in Table 1 are relative to $\gamma = 1$. External knowledge on λ can be used for calibrating the time scale, since we can expect the clearance rate to be similar in children regardless of their background. Hoti et al. (2009) and Auranen et al. (2010) estimated this rate per month to have posterior means 0.69 and 0.63 with 95% credibility intervals [0.64, 0.75] and [0.51, 0.79], respectively.

4.2. *Competition Between the Strains*

Previous longitudinal studies have assessed the competition between pneumococcal strains. Because of the biological interpretation of the parameter θ, we would expect the estimates of it to be similar throughout the studies. In Auranen et al. (2010), the posterior distribution of the competition parameter θ was estimated to have posterior mean 0.09 and a 95%-credibility interval [0.05, 0.15]. This coincides exceptionally well with our results, obtained using a different dataset and different inferential methods, and this gives cross-validatory support for all our results, as well as the conclusion on the strong inhibition of colonization, and θ having value close to 0.09.

Most studies on pneumococcal inter-strain competition have estimated competition by parameterizing the model without considering simultaneous carriage at all. Instead, models are parameterized with the *susceptibility level* parameter d, defined as the ability of a serotype to persist colonizing when a certain infection hazard of another serotype is present. Then $1 - d$ describes the relative rate to switch the carried strain to another strain. The likelihood of d is then

Table 1
Summaries of the posterior distribution of the estimated parameters, with two different simulation times for the transmission model

	Mean	Mean	95% CI	95% CI
	$T = 10$	$T = 20$	$T = 10$	$T = 20$
β	3.589	3.594	(2.8157, 4.5785)	(2.8113, 4.5621)
Λ	0.593	0.584	(0.4017, 0.8359)	(0.3875, 0.8407)
θ	0.097	0.097	(0.0605, 0.1422)	(0.0604, 0.1427)

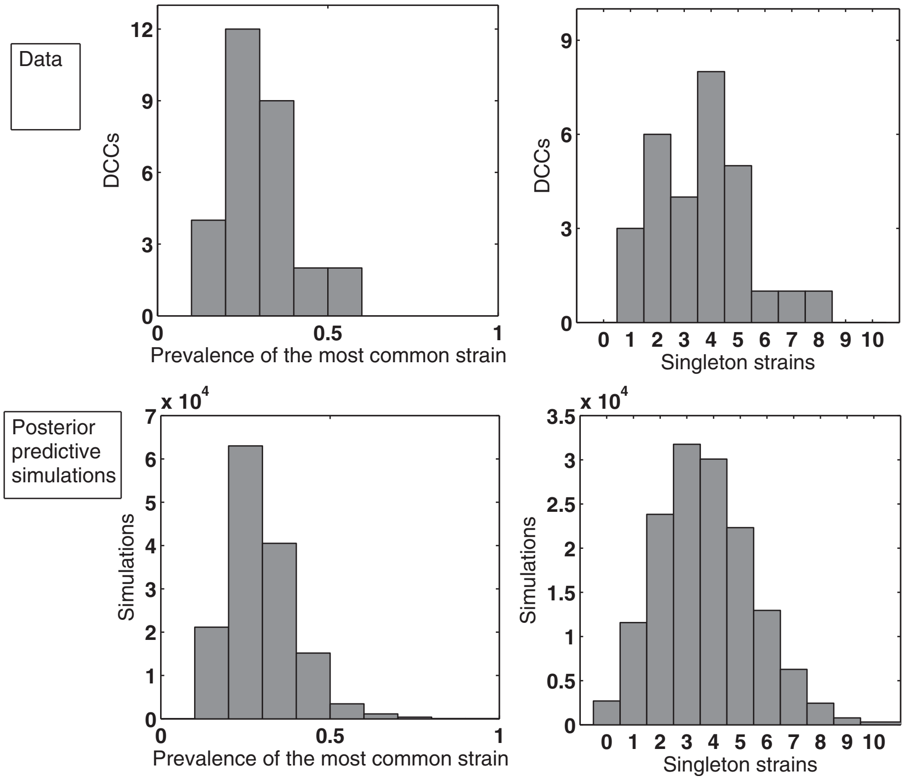

Figure 3. In the first row we show the distributions of the within-DCC measures of strain diversity as observed in the data, while in the second row we show the distributions for the same summaries predicted by the posterior predictive simulations.

constructed in terms of consequent longitudinal observations on hosts either persisting in carrying the serotype they had or switching to carry another serotype than they carried before, given the current hazards of other serotypes.

While we refer to Auranen et al. (2010) again to note that such rapid outcompetition of another serotype in the nasopharynx is not supported by data, we also note that learning about the susceptibility levels would require very dense sampling of the index individual, so that it would be possible to distinguish between consequent clearance and rapid re-infection in the index individual from an event where one serotype actually outcompetes the other. It is not clear how d is biased because the model does not account for simultaneous carriage and because data is sampled at discrete time intervals. Still, as an example, Melegaro et al. (2007) gave a point estimate for $1-d$ for the serotype 19F to be 0.26, with a 90% confidence interval [0, 0.75], Lipsitch et al. (2012) estimated it to have point estimate 0.48, with 95% confidence interval [0.37, 0.63] and Hoti et al. (2009) estimated the $1-d$, for all the serotypes, to have posterior mean 0.68 and a 95%-credibility interval [0.35, 1.10]. It is worth noticing that Lipsitch et al. (2012) assumed the serotype-specific infection hazards the individuals experience to be constant in time and the same for all individuals. As we show here, transmission is very local, and the infection hazards that the individuals experience are to a large extent determined by the strains that are carried in the local population within which most of the transmission occurs. On the other hand, in the inference framework of Hoti et al. (2009), the local strain-specific infection hazards were taken into account, but the prevalence of pneumococci was low in the study population, reported to be less than 30%. This automatically reduces the number of competition events in the population and therefore the data might be weakly informative on the absolute strength of the competition.

4.3. *Model Validation*

To further test the adequacy of our analysis, we perform posterior predictive checks by simulating 5000 realizations of $\boldsymbol{D}' \sim \Phi(\boldsymbol{D}, \boldsymbol{\varphi})$ with $\boldsymbol{\varphi}$ sampled from the posterior and simulation times $T = 10$, and use the simulation results to construct a predictive distribution for different summaries than those used in the model fitting. In particular, we consider the distributions of the observed proportion of children colonized with the most common strain in each DCC, and how many strains were observed colonizing only one host in each of the DCC's, that is, singleton strains, and assess whether the predictions coincide with the data. This idea is similar to that of Ratmann et al. (2009), who suggested consideration of posterior predictive distributions for features of data that were not used in model fitting for ABC model validation purposes, since this can reveal possible discrepancies between the model and the data. Clearly, the two summaries we use here are not independent from the summaries that were actually used in fitting. However, the degree of sufficiency of the combination of summaries we used for fitting was a priori unknown, and thus the approach makes sense. In the posterior predictive distribution, the mean for the proportion of individuals colonized with the most common strain is 0.3030 and in the data it is 0.3116. Moreover, the posterior mean for the number of singleton strains observed 3.7, while in the data it is 3.6. Thus the predictions are highly similar to the real data and also they are also very distinct from the means obtained from the 5000 simulation results with parameters sampled from the prior distribution, that were 0.2085 and 4.2904, respectively. By further examining the distributions for these summaries, as shown in Figure 3, one can see that the posterior predictive distributions of these two features are in very precise agreement with the distributions observed in the real data.

When considering the distances between the cumulative distribution functions separately, we neglected the correlations between the summary statistics used in the model fitting. Therefore, a further model validation argument is provided by the similarity of the predicted correlations to the summaries and the correlations in the data. In Web Figure 3, we also show the predicted and observed correlation structures between the two most interesting summary pairs: the Shannon diversity index and the number of observed strains, which were correlated in the data, and the number of observed strains and the prevalence of co-infection, which were not correlated in the original data.

4.4. *Model Predictions*

A very useful feature of Bayesian inference is the ability to provide statistical predictions where the uncertainty in the model parameters is coherently taken into account (Bernardo and Smith, 1994). To assess posterior predictive distributions, we utilized the results of 5000 simulations with parameters sampled from the posterior, and let the simulation time exceed $T = 10$ before assessing the predicted feature. Firstly, we constructed the posterior predictive distribution of the prevalence of colonization and co-infection among *all* the 53 individuals in the DCC, as shown in Web Figure 4. The posterior predictive distribution of the true prevalence of colonization within 53 individuals in a single DCC had mean 0.794. with 95% credibility interval [0.660, 0.906]. Notice that the prevalence of colonization *among the studied individuals* was higher than 0.9 in 5 of the studied DCCs, as seen from Figure 1, but our predictions suggest that it is improbable that the total prevalences were that high in all of these DCCs. Finally, the posterior predictive prevalence of co-infection among the colonized individuals in a DCC had mean 0.133, with the confidence interval [0.026, 0.256].

Since we estimated the parameters relative to the clearance rate, to assess the model predictions in calendar time, we need to calibrate the time scale. For this purpose, we use 0.69 as the estimate for clearance rate *per month*, as this value was attained a high posterior probability in the two longitudinal studies of Auranen et al. (2010) and Hoti et al. (2009). Given this time scale, according to the predictive simulations, on average 5.7 new outbreaks are introduced to a DCC from the community during one month. A total of 54.1% of these outbreaks are predicted to be of size one, meaning that the host infected from the community does not cause any further infections in the DCC. Furthermore, the mean outbreak size is 6.3, and 95% of the outbreaks are predicted to end up being smaller than 27. The posterior predictive distributions describing the properties of the epidemics within the DCC are shown in more detail in Figure 4.

Figure 4 also shows the empirical distribution function of the posterior predictive distribution of the time to first colonization event for a susceptible individual, and also the proportion of time spent colonized for an individual attending a DCC for one year. We predict that the mean time for a susceptible individual to become colonized in a DCC is 0.4346 months, and 0.95% of initially susceptible individuals have been colonized at least once, after spending 1.2625 months in a DCC. On average, an initially susceptible individual spends a proportion 0.6832 (95% credibility interval [0.254, 0.968]) colonized. We conclude that while the prevalence of colonization is high, there exists a considerable variation in the proportion of time that an individual actually spends colonized, due to being fortunate in escaping colonization or the opposite.

5. Discussion

Here we have considered a transmission model under which prevalences of different strains in local populations were determined by both the global prevalences of strains, and also the interplay of the three types of stochastic event occurring in the DCC: transmissions from the community, that introduce new strains to DCC; transmissions within the DCC, that amplify the prevalence of a particular strain; and the clearance of colonization, that ensures turnover of the local strain population. This model was then fitted to a cross-sectional dataset of observed prevalences of different strains in 29 different DCCs. We assumed that the snapshots of local strain prevalences in data were realizations from the stationary distribution of such transmission processes, and showed that the between-DCC variations in strain diversity, colonization and co-infection, together reveal the characteristics of the processes described. As the local diversity patterns of strains were considered as a basis for model fitting, our approach can be interpreted as an extension of approaches based on final-size distributions of epidemics. Traditionally, final-size approaches consider epidemics involving pathogens, for which immunity can be obtained, and thus it is reasonable to consider the actual final sizes of epidemics. Local prevalences of strains are also informative about the epidemic spread within a DCC, because they reflect the distribution of sizes of microepidemics of different strains, which in turn are determined by the processes described above.

The framework presented here has a few assumptions that should be kept in mind. Assuming both the global prevalence of colonization and the global strain distribution to be stable might seem vague. However, follow-up studies monitoring the effects of conjugate vaccines to pneumococcal populations have all observed that the prevalence of colonization remained at its past level a few years after the introduction of vaccines. This phenomenon was also observed in the study population here, as shown by a follow-up study, where data was collected in similar manner as the data analyzed here (Vestrheim et al., 2010). The fact that the prevalence of colonization retains the same level after a large scale ecological perturbation to the strain community suggests that the community-level

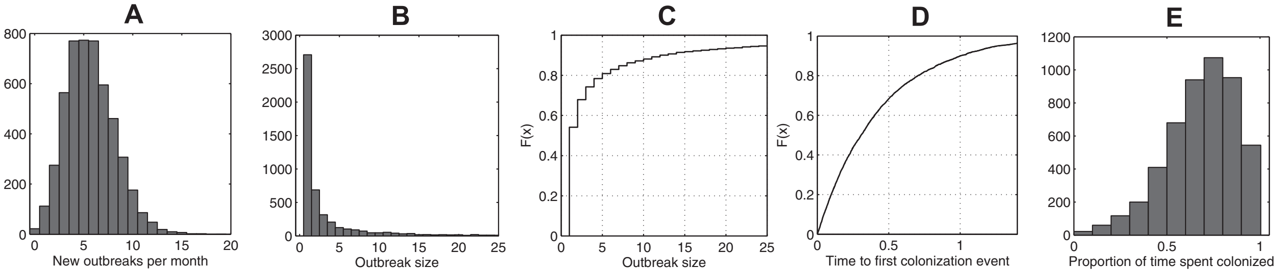

Figure 4. (A) Posterior predictive distribution of the number of epidemics initiated in the DCC from the community during 1 month and (B) and (C) Sizes of the epidemics, showing that the majority of infections from the community do not lead to any further transmissions within the DCC. (D) Posterior predictive distribution for the time in months until a susceptible individual is colonized for the first time in a DCC, and (E) Distribution of the proportion of time that initially a susceptible DCC attendee in total is colonized in total during 1 year.

prevalence of colonization can be considered to be very constant. Similarly, it was pointed out by Hanage (2010) that seven years after the launch of conjugate vaccine usage in Massachusetts, the strain distribution exhibited a similar population structure and similar levels of diversity as before the vaccines: only the strain composition was changed. This suggests that the strain distribution in a large community can be also considered to be very stable.

Our framework also assumed that the observations in the data were from stationary distribution of the transmission process. Because the rate of convergence to the stationary distribution is different for different model parameters, we emphasize the importance of exploratory analysis to determine the robustness of the simulation results to the simulation time. Also, the framework presented could be adjusted for a situation where the assumption of the DCCs being of identical sizes should be relaxed. The simulations can also be performed for different population sizes, only the discrepancy measure for ABC should be carefully adjusted, since the simulations are no longer realizations from identical processes. The fact that we simulated observations from DCCs of equal sizes which might yield slightly over-confident parameter estimates, if the true sizes did vary considerably. However, there is a fair amount of support for the assumptions and the results obtained, both in light of the earlier longitudinal study by Auranen et al. (2010) and the posterior predictive checks; the obtained posterior distribution for model parameters was demonstrated to capture the system behavior with a high level of accuracy.

The posterior distributions of model parameters are useful for formulating predictive statements of the system behavior, as illustrated in the previous section. Based on such knowledge, it is also possible to predict how the system would respond to environmental changes, for example those related to control strategies, such as vaccines or antibiotics. Furthermore, understanding the transmission dynamics allows the assessment of different evolutionary scenarios. Fraser, Hanage, and Spratt (2005) suggested that a neutral model for the evolution of the pneumococcal genome is adequate to describe the MLST data of the pneumococcus, once the model takes into account the microepidemic spread of the pneumococcal strains. Our approach disentangled the composition of micro-epidemics in the study population, and information of this type could be used to assess the hypothesis of the neutral evolution of pneumococcus even further. It is also worth emphasizing the good fit of our model which assumed equality of strains in their ecological properties. We are aware that Weinberger et al. (2009) detected differences between the performance of pneumococcal strains with laboratory experiments in mice, especially in the duration of carriage, but it is possible that these differences are negligible when considering the dynamics of pneumococcal strains in actual human populations. On the other hand, as pointed out by Cobey and Lipsitch (2012), as individuals age, the duration of carriage declines and approaches a fixed value, which is the same for all of the serotypes, and this stabilizes competition between the strains slightly. This, together with the rare strains having more infected individuals to co-colonize than the common strains, is a possible explanation for the persistence of the less fit strains and neutral-alike observations. The question of neutrality could be assessed more carefully in the future by assessing whether observations on the vaccine-induced serotype replacement coincide with neutral expectations. On the other hand, to assess the actual differences between the strains would require detailed data with sufficient observations on rare strains. Such information is obviously useful, for example when deciding the target strains of conjugate vaccines.

Our results lend support for the strong inhibition of colonization as a mechanism of between-strain competition. As hypothesized by Donkor et al. (2011) the co-occurrence of different strains in the nasopharynx is of interest when assessing the evolution and genetic diversity of the pneumococcus, as it is known from previous studies that the pneumococcus is a highly recombining bacterium (Feil et al., 2000). Moreover, this is all related to the questions surrounding bacterial speciation (Fraser, Hanage, and Spratt, 2007), that is, the appearance of clusters of bacteria that are genetically similar, and the role of recombination in that process, as the probability of recombination between different pneumococcal strains is restricted by the occasion of co-infection. Competition between strains is also of interest when evaluating vaccination strategies. In previous studies, competition between strains has been addressed in several ways, but the results in different studies are contradictory, and model structures also contradict what is known about pneumococcal competition. Moreover, due to the clustering of infectious contacts in local populations, in order to assess the competition, the local serotype-specific hazards that change in time should be assessed realistically, as we did here.

Recruiting individuals for longitudinal studies, where swabs are typically taken monthly, is more difficult than performing a cross-sectional study, where a single sample per individual suffices. Thus, the relevancy of our approach to questions of statistical study design is also worth considering. For instance, one can use pseudo-observed datasets simulated with known parameter values for testing the robustness of the methods to identify the true parameter values. Then the minimum requirements for the data to be sufficiently informative about the model parameters can be identified. Hence a similar approach could be utilized for studying the epidemiology of other pathogens for which there are several strains co-circulating, such as *Neisseria meningitidis* or *Mycobacterium tuberculosis*.

6. Supplementary Material

The web figures referenced in Sections 2.1, 3.2, 4.3, and 4.4, together with a detailed description of the sequential ABC-procedure that was performed, are available with this paper at the Biometrics website on Wiley Online Library. The matlab codes for performing similar computations are also provided there.

Acknowledgements

We would like to thank Didrik Vestrheim for kindly providing access to the pneumococcal data set. Also, for useful comments, we thank the anonymous referees, Mikhail Shubin, Marc Lipsitch, William Hanage and the participants of the third Permafrost workshop. This research was supported by ERC grant no. 239784, Academy of Finland grant no. 251170 and the FICS and Finnish Population Genetics Graduate Schools.

References

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control.* Oxford: Oxford University Press. 757 p.

Auranen, K., Mehtälä, J., Tanskanen, A., and Kaltoft, M. S. (2010). Between-strain competition in acquisition and clearance of pneumococcal carriage—Epidemiologic evidence from a longitudinal study of day-care children. *American Journal of Epidemiology* **171**, 169–176.

Ball, F., Mollison, D., and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *Annals of Applied Probability* **7**, 46–89.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics* **41**, 379–406.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory. Wiley Series in Probability and Statistics.* Chichester: John Wiley & Sons 586 p.

Bogaert, D., De Groot, R., and Hermans, P. W. (2004). *Streptoccus pneumoniae* colonisation: The key to pneumococcal disease. *The Lancet Infectious Diseases* **4**, 144–54.

Brooks-Pollock, E., Becerra, M. C., Goldstein, E., Cohen, T., and Murray, M. B. (2011). Epidemiologic inference from the distribution of tuberculosis cases in households in Lima, Peru. *The Journal of Infectious Diseases* **203**, 1582–9.

Cobey, S. and Lipsitch, M. (2012). Niche and neutral effects of acquired immunity permit coexistence of pneumococcal serotypes. *Science* **335**, 1376.

Donkor, E. S., Bishop, C. J., Antonio, M., Wren, B., and Hanage, W. P. (2011). High levels of recombination among *Streptoccus pneumoniae* isolates from the Gambia. *mBio* **2**, e00040-11.

Dunais, B., Pradier, C., Carsenti, H., Sabah, M., Mancini, G., Fontas, E., and Dellamonica, P. (2004). Influence of child care on nasopharyngeal carriage of *Streptoccus pneumoniae* and *Haemophilus influenzae*. *The Pediatric Infectious Disease Journal* **22**, 589–92.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics* **22**, 403–434.

Feil, E. J., Smith, J. M., Enright, M. C., and Spratt, B. G. (2000). Estimating recombinational parameters in *Streptoccus pneumoniae* from multilocus sequence typing data. *Genetics* **154**, 1439–50.

Fraser, C., Hanage, W. P., and Spratt, B. G. (2005). Neutral microepidemic evolution of bacterial pathogens. *Proceedings of the National Academy of Sciences* **102**, 1968–1973.

Fraser, C., Hanage, W. P., and Spratt, B. G. (2007). Recombination and the nature of bacterial speciation.. *Science* **315**, 476–480.

Hagenaars, T. J., Donnelly, C. A., and Ferguson, N. M. (2004). Spatial heterogeneity and the persistence of infectious diseases. *Journal of Theoretical Biology* **229**, 349–359.

Hanage, W. P. (2010). Evidence that pneumococcal serotype replacement in Massachusetts following conjugate vaccination is now complete. *Epidemics* **2**, 80–84.

Hoti, F., Erästö, P., Leino, T., and Auranen, K. (2009). Outbreaks of *Streptoccus pneumoniae* carriage in day care cohorts in Finland—Implications for elimination of transmission. *BMC Infectious Diseases* **9**, 102.

Huang, S. S., Finkelstein, J. A., and Lipsitch, M. (2009). Modeling community-and individual-level effects of childcare center attendance on pneumococcal carriage. *Clinical Infectious Diseases* **40**, 1215–22.

Hubbel, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography.* Princeton: Princeton University Press. 448 p.

Kellner, J. D., Scheifele, D., Vanderkooi, O. G., MacDonald, J., Church, D. L., and Tyrrel, G. J. (2008). Effects of routine infant vaccination with the 7-valent pneumococcal conjugate vaccine on nasopharyngeal colonization with *Streptoccus pneumoniae* in children in Calgary, Canada. *The Pediatric Infectious Disease Journal* **27**, 526–32.

Lipsitch, M., Abdullahi, O., D'Amour, A., Wen, X., Weienberger, D. M., Tchechen, E., and Scott, J. A. G. (2012). Estimating rates of carriage acquisition and clearance and competitive ability for pneumococcal serotypes in kenya with a Markov transition model. *Epidemiology* **23**, 510–9.

Luciani, F., Sisson, S. A., Jiang, H., Francis, A. R., and Tanaka, M. M. (2009). The epidemiological fitness cost of drug resistance in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences* **106**, 14711–14715.

Melegaro, A., Choi, Y., Pebody, R., and Gay, N. (2007). Pneumococcal carriage in United Kingdom families: Estimating serotype-specific transmission parameters from longitudinal data. *American Journal of Epidemiology* **166**, 228–235.

Peet, R. K. (1974). The measurement of species diversity, *Annual Review of Ecology and Systematics* **5**, 285–307.

Pritchard, J., Seielstad, M., Perez-Lezaun, A., and Feldman, M. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16**, 1791–1798.

Ratmann, O., Andrieu, C., Wiuf, C., and Richardson, S. (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences* **106**, 10576–10581.

Sisson, S. S., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **104**, 1760–1765.

Stone, P., Wilkinson-Herbots, H., and Isham, V. (2008). A stochastic model for head lice infections. *Journal of Mathematical Biology* **56**, 746–764.

Vestrheim, D. F., Hiby, E. A., Aaberge, I. S., and Caugant, D. A. (2008). Phenotypic and genotypic characterization of *Streptoccus pneumoniae* strains colonizing children attending day-care centers in Norway. *Journal of Clinical Microbiology* **46**, 2508–2518.

Vestrheim, D. F., Hiby, E. A., Aaberge, I. S., and Caugant, D. A. (2010). Impact of a pneumococcal conjugate vaccination program on carriage among children in Norway. *Clinical and Vaccine Immunology* **17**, 325–334.

Weinberger, D. M., Trzcinski, K., Lu, Y. J., Bogaert, D., Brandes, A., Galagan, A., Anderson, P. W., Malley, R., and Lipsitch, M. (2009). Pneumococcal capsular polysaccharide structure predicts serotype prevalence. *PLOS Pathogens* **5**, e1000476.

Received May 2012. Revised February 2013.
Accepted February 2013.