eAppendices

January 2, 2013

1 eAppendix A: Statistical methods

We consider a model for n strains assuming that an individual can be colonized with up to two strains simultaneously. In a small number of samples in each of the three datasets, simultaneous colonization of three serotypes was found. In the analysis, the isolate with the least observations among all isolates in the respective dataset was then omitted. The possible states of an individual thus are: non-colonized, colonized with one of the n strains, and colonized with any two of the strains simultaneously. We label the states as $(0), (1), ..., (n), (1,2), (1,3), ..., (n-1,n)$, where (0) corresponds to the non-colonized state, states $(1), ..., (n)$ to the colonized states by one strain, and states $(1,2), ..., (n-1,n)$ to the states in which two strains simultaneously colonize the same individual. Altogether, there are $N_n = (n(n+1)/2)+1$ states in the model. Occasionally, it is more convenient to enumerate the states as $[1], ..., [N_n]$, where $[1]$ corresponds to state (0) etc.

Denote the process of colonization by Y. A $N_n \times N_n$ stochastic transition intensity matrix $\mathbf{M}(\tau)$ at time τ is defined element-wise as

$$\lambda_{[s],[k]} = \lim_{\Delta t \to 0} \frac{\mathbb{P}(Y(\tau + \Delta t) = [k] \mid Y(\tau) = [s])}{\Delta t},$$

where $[s], [k] \in \{[1], ..., [N_n]\}$, $s \neq k$, and

$$\lambda_{[s],[s]} = - \sum_{j=1, j \neq s}^{N_n} \lambda_{[s],[j]}.$$

With constant transition intensities ($\mathbf{M}(\tau) = \mathbf{M}$) the transition probability matrix for the time interval of length t is obtained with the matrix exponential function $\mathbb{P}^t(\mathbf{M}) = \exp(t\mathbf{M})$, with elements

$$\mathbb{P}^t_{[s],[k]}(\mathbf{M}) = \mathbb{P}(Y(\tau + t) = [k] \mid Y(\tau) = [s]) \text{ for all } \tau.$$

Assume that the status of individual i is observed at time points $\tau_i^j,\ j = 1, ..., K_i$. Based on the observations from N individuals, the likelihood function for the complete data $Y = \{Y(\tau_i^j); i = 1, ..., N, j = 1, ..., K_i\}$ is

$$\prod_{i=1}^{N} \mathbb{P}(Y(\tau_i^1)|\mathbf{M}_0) \prod_{j=2}^{K_i} \mathbb{P}(Y(\tau_i^j)|Y(\tau_i^{j-1}), \mathbf{M}),$$

where $\mathbf{M}_0$ is the distribution of the first observation $Y(\tau_i^1)$. Without the contribution of the first observations, the complete-data likelihood function is

$$\prod_{i=1}^{N}\prod_{j=2}^{K_i}\sum_{s=1}^{N_n}\sum_{k=1}^{N_n} \mathbb{P}^{t(i,j)}_{[s],[k]}(\mathbf{M})\mathbf{1}\{Y(\tau_i^j) = [k], Y(\tau_i^{j-1}) = [s]\},$$

where $t(i,j) = \tau_i^j - \tau_i^{j-1}$. If the true underlying states of the individuals are different from what is observed, this likelihood is not applicable. To allow such possibility, we present the necessary methods in the following.

A hidden Markov model approach. To allow the possibility that the true state of colonization is not necessarily observed, we need to specify a model to link the observations to the true underlying process. For the non-colonized state and for any of the n singly colonized states we assume perfect sensitivity of detection, i.e., any of the states $(0), ..., (n)$ is observed as such when measured. For any of the doubly colonized states, $(1, 2), ..., (n-1, n)$, we assume that the state is detected as either the true state with two strains, or a singly colonized state with one of the two strains involved. We define $0 \leq \nu \leq 1$ as the sensitivity to detect the doubly colonized state as such. A doubly colonized state is then detected as singly colonized with probability $1 - \nu$ so that either of the two strains is detected with probability $0.5(1 - \nu)$.

Denote the discrete-time observations made at times τ_i^j by $X = \{X(\tau_i^j), i = 1, ..., N, j = 1, ..., K_i\}$. The likelihood function for the incomplete data X is

$$\prod_{i=1}^{N} \prod_{j=2}^{K_i} \mathbb{P}(X(\tau_i^j)|\mathcal{F}(X(\tau_i^{j-1}), \mathbf{M}), \tag{1}$$

where $\mathcal{F}(X(\tau_i^{j-1}))$ is the observed history of colonization states of individual i up to time point τ_i^{j-1}. This model is a hidden Markov model because of the Markov dependency between the underlying states $(Y(\tau_i^j))$. To estimate the transition intensities $\mathbf{M}$ we modify the algorithm originally presented by Nagelkerke[1] for a simple binary model to the current setting with more than two states. The resulting algorithm is similar to the Viterbi algorithm.[2]

Denote the probability that the true state of individual i at time τ_i^j is $y \in S = \{(0), ..., (n-1, n)\}$, conditional on the observed history, by

$$Q_i^j(y) = \mathbb{P}(Y(\tau_i^j) = y|\mathcal{F}(X(\tau_i^j)), \mathbf{M}).$$

In the following, we leave conditioning on $\mathbf{M}$ and the time interval $t(i,j) = \tau_i^j - \tau_i^{j-1}$ from the notations, and denote the transition probability from state (s) to (k), i.e., the element $\mathbb{P}_{(s),(k)}^{t(i,j)}(\mathbf{M})$, by $p((s),(k))$. For $y \in \{(1,2), ..., (n-1,n)\}$, the states with simultaneous colonization of two strains, the probability of observation $X(\tau_i^j)$, conditional on the observed history $\mathcal{F}(X(\tau_i^{j-1}))$, is

$$\mathbb{P}(X(\tau_i^j) = y|\mathcal{F}(X(\tau_i^{j-1})) = \nu \sum_{k \in S} \left[Q_i^{j-1}(k) p(k,y) \right],$$

For $y \in \{(1), ..., (n)\}$, the states in which only one colonizing strain is present,

$$\mathbb{P}(X(\tau_i^j) = y|\mathcal{F}(X(\tau_i^{j-1})) = \sum_{k \in S} \left[Q_i^{j-1}(k) p(k,y) + 0.5(1-\nu) \sum_{d \in \{(y,\cdot)\}} Q_i^{j-1}(k) p(k,d) \right],$$

where $\{(y, \cdot)\}$ is the set of states involving two simultaneous strains where the other one is y. For (0), the non-colonized state,

$$\mathbb{P}(X(\tau_i^j) = (0)|\mathcal{F}(X(\tau_i^{j-1})) = \sum_{k \in S} \left[Q_i^{j-1}(k) p(k,(0)) \right].$$

Next we use the Bayes theorem to obtain an expression for $Q_i^j(y)$:

$$
\begin{aligned}
Q_i^j(y) &= \frac{\mathbb{P}\Bigg(\mathcal{F}(X(\tau_i^j)), Y(\tau_i^j) = y\Bigg)}{\sum_{s\in S}\Bigg[\mathbb{P}\Bigg(\mathcal{F}(X(\tau_i^j)), Y(\tau_i^j) = s\Bigg)\Bigg]} \\
&= \frac{\mathbb{P}\Bigg(X(\tau_i^j)|Y(\tau_i^j)=y, \mathcal{F}(X(\tau_i^{j-1}))\Bigg)\mathbb{P}\Bigg(Y(\tau_i^j)=y|\mathcal{F}(X(\tau_i^{j-1}))\Bigg)\mathbb{P}\Bigg(\mathcal{F}(X(\tau_i^{j-1}))\Bigg)}{\sum_{s\in S}\Bigg[\mathbb{P}\Bigg(X(\tau_i^j)|Y(\tau_i^j)=s, \mathcal{F}(X(\tau_i^{j-1}))\Bigg)\mathbb{P}\Bigg(Y(\tau_i^j)=s|\mathcal{F}(X(\tau_i^{j-1}))\Bigg)\mathbb{P}\Bigg(\mathcal{F}(X(\tau_i^{j-1}))\Bigg)\Bigg]} \\
&= \frac{\mathbb{P}\Bigg(X(\tau_i^j)|Y(\tau_i^j)=y, \mathcal{F}(X(\tau_i^{j-1}))\Bigg)\mathbb{P}\Bigg(Y(\tau_i^j)=y|\mathcal{F}(X(\tau_i^{j-1}))\Bigg)}{\sum_{s\in S}\Bigg[\mathbb{P}\Bigg(X(\tau_i^j)|Y(\tau_i^j)=s, \mathcal{F}(X(\tau_i^{j-1}))\Bigg)\mathbb{P}\Bigg(Y(\tau_i^j)=s|\mathcal{F}(X(\tau_i^{j-1}))\Bigg)\Bigg]}. \quad (2)
\end{aligned}
$$

The observation level in (2), $\mathbb{P}(X(\tau_i^j)|Y(\tau_i^j), \mathcal{F}(X(\tau_i^{j-1})))$, depends only on ν and thus follows from the model linking observations to the underlying process. Given observations up to time point τ_i^{j-1}, the probability that the state at τ_i^j is y can be calculated as follows:

$$
\begin{aligned}
&\mathbb{P}(Y(\tau_i^j) = y|\mathcal{F}(X(\tau_i^{j-1}))) = \sum_{s\in S}\Bigg[\mathbb{P}(Y(\tau_i^j) = y, Y(\tau_i^{j-1}) = s|\mathcal{F}(X(\tau_i^{j-1}))\Bigg] \\
&= \sum_{s\in S}\Bigg[\mathbb{P}(Y(\tau_i^j) = y \mid Y(\tau_i^{j-1}) = s, \mathcal{F}(X(\tau_i^{j-1}))\mathbb{P}(Y(\tau_i^{j-1}) = s|\mathcal{F}(X(\tau_i^{j-1}))\Bigg] \\
&= \sum_{s\in S}\Bigg[\mathbb{P}(Y(\tau_i^j) = y \mid Y(\tau_i^{j-1}) = s, \mathcal{F}(X(\tau_i^{j-1}))Q_i^{j-1}(s)\Bigg] \\
&= \sum_{s\in S}\Bigg[p(s,y)Q_i^{j-1}(s)\Bigg],
\end{aligned}
$$

yielding a recursive formula to calculate the transition probabilities needed in (1). The recursion stops when the observation determines the underlying state (either 0 or doubly colonized state) or on the first state, which for simplicity is assumed to be the observed one.

We implemented a Markov chain Monte Carlo method with the Metropolis-Hastings algorithm to draw samples from the posterior distribution of all $(2n+6)$ parameters. We assumed uniform prior distributions on $[0, 10]$ (per month) for all acquisition and clearance rates ($\lambda_{(0),(x)}$ and $\lambda_{(x),(0)}$). For all the competition parameters ($\theta_{(x),(x,y)} = \lambda_{(x),(x,y)}/\lambda_{(0),(y)}$ and $\phi_{(x,y),(x)} = \lambda_{(x,y),(x)}/\lambda_{(y),(0)}$), we assumed prior distributions proportional to z^{-1} on $z \in [1/a, a], 1 < a$ ($a = 1000$), since this prior has the property that $[1/b, 1]$ and $[1, b]$ have the same probability for all $1 < b \leq a$. We monitored the convergence of the chain by setting different initial values for the model parameters to start the chain and calculated the potential scale reduction factor[3] for the parameters of interest. The detection sensitivity, ν, was not subject to estimation but used as a control parameter with a given value. The performance of estimation was investigated with different choices of ν. Specifically, values $\nu = 50\%$ and $\nu = 100\%$ were used when estimating the model parameters.

Interpretation of the overall competition strength $(1 - \theta/\phi)$**.** Consider one time unit spent singly colonized in absence of competition and denote by λ the rate of acquiring double colonization while singly colonized. During the time unit, on average λ double colonization events will occur (expectation of the Poisson distribution). Denote by μ the rate of clearing double colonization, so that each double colonization episode has an average duration $1/\mu$. Then, per one time unit spent singly colonized the expected time spent double colonized is λ/μ. Suppose then that in presence of competition the corresponding acquisition and clearance rates are $\theta\lambda$ and $\varphi\mu$, respectively. Then, per one time unit spent singly colonized, the expected time spent doubly colonized is $(\theta\lambda)/(\varphi\mu)$. This means that the relative reduction in the expected time spent doubly colonized per time unit spent singly colonized is $1 - \theta/\varphi$, compared to no competition.

Simulation studies. Simulated data were used to test the performance of estimation. In these simulations, 8 equivalent strains with acquisition rates 0.15 per month, and clearance rate 0.75 per month were used. In addition, the competition parameters (θ, ϕ) were one of the following combinations: (0.1,1),(1,1),(1,5). These correspond to the overall and double colonization prevalence as (0.63,0.07),(0.73,0.41),(0.65,0.13), respectively. The average durations of doubly colonized state were 0.67,0.67,0.13 months, respectively. Four different combinations of the sensitivity to simulate the data (ν_{true}) and the sensitivity used to estimate the model parameters (ν_{est}) were used: (0.25,1), (0.5,1), (0.5,0.5), (1,0.5). The two first combinations correspond to analyzes in which perfect sensitivity is assumed although the true sensitivity to detect double colonization is poor. The last combination represents the case in which the true sensitivity is higher than the one used in estimation.

2 eAppendix B: Estimation results

Table 1: The Danish dataset

$\boldsymbol{\nu = 100\%}$	**Mean**	**90%CI**		**Mean**	**90%CI**
$\lambda_{(0),(23\text{F})}$	0.54	$(0.37, 0.76)$	$\lambda_{(23\text{F}),(0)}$	1.00	$(0.58, 1.41)$
$\lambda_{(0),(19\text{F})}$	0.59	$(0.35, 0.96)$	$\lambda_{(19\text{F}),(0)}$	1.28	$(0.81, 2.08)$
$\lambda_{(0),(6\text{A})}$	0.17	$(0.10, 0.26)$	$\lambda_{(6\text{A}),(0)}$	0.67	$(0.42, 0.98)$
$\lambda_{(0),(14)}$	0.23	$(0.14, 0.36)$	$\lambda_{(14),(0)}$	1.28	$(0.81, 1.94)$
$\lambda_{(0),(6\text{B})}$	0.32	$(0.21, 0.47)$	$\lambda_{(6\text{B}),(0)}$	0.93	$(0.61, 1.29)$
$\lambda_{(0),(19\text{A})}$	0.13	$(0.07, 0.22)$	$\lambda_{(19\text{A}),(0)}$	0.96	$(0.57, 1.51)$
$\lambda_{(0),(15\text{B/C})}$	0.23	$(0.10, 0.42)$	$\lambda_{(15\text{B/C}),(0)}$	1.61	$(0.71, 2.77)$
$\lambda_{(0),(11\text{A})}$	0.28	$(0.14, 0.49)$	$\lambda_{(11\text{A}),(0)}$	2.51	$(1.44, 3.73)$
$\lambda_{(0),(\text{The Rest})}$	0.79	$(0.48, 1.19)$	$\lambda_{(\text{The Rest}),(0)}$	1.28	$(0.79, 1.91)$
$\theta_{(x),(x,23\text{F})}$	0.04	$(0.01, 0.14)$	$\phi_{(23\text{F},x),(x)}$	1.09	$(0.13, 2.90)$
$\theta_{(23\text{F}),(23\text{F},x)}$	0.10	$(0.02, 0.21)$	$\phi_{(x,23\text{F}),(23\text{F})}$	0.61	$(0.02, 1.70)$
$\theta_{(x),(x,y)}$	0.10	$(0.04, 0.21)$	$\phi_{(x,y),(x)}$	0.94	$(0.49, 2.11)$
$\boldsymbol{\nu = 50\%}$	**Mean**	**90%CI**		**Mean**	**90%CI**
$\lambda_{(0),(23\text{F})}$	0.53	$(0.30, 0.83)$	$\lambda_{(23\text{F}),(0)}$	1.28	$(0.74, 1.92)$
$\lambda_{(0),(19\text{F})}$	0.48	$(0.29, 0.75)$	$\lambda_{(19\text{F}),(0)}$	1.15	$(0.68, 1.87)$
$\lambda_{(0),(6\text{A})}$	0.16	$(0.09, 0.24)$	$\lambda_{(6\text{A}),(0)}$	0.70	$(0.45, 1.05)$
$\lambda_{(0),(14)}$	0.25	$(0.14, 0.40)$	$\lambda_{(14),(0)}$	1.62	$(0.91, 2.43)$
$\lambda_{(0),(6\text{B})}$	0.34	$(0.21, 0.54)$	$\lambda_{(6\text{B}),(0)}$	1.12	$(0.70, 1.69)$
$\lambda_{(0),(19\text{A})}$	0.13	$(0.07, 0.21)$	$\lambda_{(19\text{A}),(0)}$	1.08	$(0.63, 1.67)$
$\lambda_{(0),(15\text{B/C})}$	0.17	$(0.09, 0.29)$	$\lambda_{(15\text{B/C}),(0)}$	1.40	$(0.69, 2.33)$
$\lambda_{(0),(11\text{A})}$	0.28	$(0.13, 0.48)$	$\lambda_{(11\text{A}),(0)}$	2.74	$(1.57, 4.17)$
$\lambda_{(0),(\text{The Rest})}$	0.76	$(0.46, 1.25)$	$\lambda_{(\text{The Rest}),(0)}$	1.38	$(0.83, 2.28)$
$\theta_{(x),(x,23\text{F})}$	0.09	$(0.01, 0.26)$	$\phi_{(23\text{F},x),(x)}$	0.25	$(0.00, 0.88)$
$\theta_{(23\text{F}),(23\text{F},x)}$	0.06	$(0.00, 0.17)$	$\phi_{(x,23\text{F}),(23\text{F})}$	0.42	$(0.00, 1.00)$
$\theta_{(x),(x,y)}$	0.10	$(0.05, 0.18)$	$\phi_{(x,y),(x)}$	0.32	$(0.17, 0.58)$

Table 1. Estimates of acquisition ($\lambda_{(0,.)}$), clearance ($\lambda_{(.,0)}$) and competition parameters ($\theta_{(.,.)}, \phi_{(.,.)}$). The posterior mean and 90% credible intervals (CI; in parenthesis) with the detection sensitivity (ν) of double colonization assumed either 100% or 50%. The target serotype (s) is 23F.

Table 2: The American Indian dataset

$\boldsymbol{\nu = 100\%}$	**Mean**	**90%CI**		**Mean**	**90%CI**
$\lambda_{(0,\text{6A})}$	0.08	$(0.04, 0.14)$	$\lambda_{(\text{6A},0)}$	0.23	$(0.10, 0.41)$
$\lambda_{(0,\text{6B})}$	0.06	$(0.04, 0.09)$	$\lambda_{(\text{6B},0)}$	0.42	$(0.29, 0.55)$
$\lambda_{(0,\text{23F})}$	0.08	$(0.05, 0.11)$	$\lambda_{(\text{23F},0)}$	0.52	$(0.36, 0.71)$
$\lambda_{(0,\text{19F})}$	0.10	$(0.07, 0.15)$	$\lambda_{(\text{19F},0)}$	0.99	$(0.66, 1.41)$
$\lambda_{(0,14)}$	0.05	$(0.03, 0.07)$	$\lambda_{(14,0)}$	0.58	$(0.39, 0.81)$
$\lambda_{(0,\text{19A})}$	0.04	$(0.02, 0.06)$	$\lambda_{(\text{19A},0)}$	0.52	$(0.33, 0.77)$
$\lambda_{(0,\text{22F})}$	0.03	$(0.02, 0.05)$	$\lambda_{(\text{22F},0)}$	0.45	$(0.29, 0.63)$
$\lambda_{(0,\text{9V})}$	0.07	$(0.04, 0.11)$	$\lambda_{(\text{9V},0)}$	1.25	$(0.78, 1.85)$
$\lambda_{(0,\text{The rest})}$	0.73	$(0.57, 0.90)$	$\lambda_{(\text{The rest},0)}$	0.91	$(0.70, 1.15)$
$\theta_{(x),(x,y)}$	0.28	$(0.13, 0.54)$	$\phi_{(x,y),(x)}$	2.55	$(1.23, 4.96)$
$\boldsymbol{\nu = 50\%}$	**Mean**	**90%CI**		**Mean**	**90%CI**
$\lambda_{(0,\text{6A})}$	0.08	$(0.04, 0.12)$	$\lambda_{(\text{6A},0)}$	0.42	$(0.27, 0.57)$
$\lambda_{(0,\text{6B})}$	0.06	$(0.04, 0.08)$	$\lambda_{(\text{6B},0)}$	0.38	$(0.27, 0.50)$
$\lambda_{(0,\text{23F})}$	0.08	$(0.05, 0.11)$	$\lambda_{(\text{23F},0)}$	0.56	$(0.41, 0.79)$
$\lambda_{(0,\text{19F})}$	0.08	$(0.05, 0.13)$	$\lambda_{(\text{19F},0)}$	0.85	$(0.57, 1.18)$
$\lambda_{(0,14)}$	0.04	$(0.03, 0.07)$	$\lambda_{(14,0)}$	0.60	$(0.40, 0.80)$
$\lambda_{(0,\text{19A})}$	0.04	$(0.03, 0.06)$	$\lambda_{(\text{19A},0)}$	0.54	$(0.35, 0.76)$
$\lambda_{(0,\text{22F})}$	0.03	$(0.02, 0.05)$	$\lambda_{(\text{22F},0)}$	0.48	$(0.33, 0.64)$
$\lambda_{(0,\text{9V})}$	0.05	$(0.03, 0.09)$	$\lambda_{(\text{9V},0)}$	1.03	$(0.57, 1.57)$
$\lambda_{(0,\text{The rest})}$	0.64	$(0.55, 0.73)$	$\lambda_{(\text{The rest},0)}$	0.82	$(0.72, 0.93)$
$\theta_{(x),(x,y)}$	0.31	$(0.20, 0.45)$	$\phi_{(x,y),(x)}$	1.19	$(0.75, 1.69)$

Table 2. Estimates of acquisition ($\lambda_{(0,.)}$), clearance ($\lambda_{(.,0)}$) and competition parameters ($\theta_{(.,.)}, \phi_{(.,.)}$). The posterior mean and 90% credible intervals (CI; in parenthesis) with the detection sensitivity (ν) of double colonization assumed

either 100% or 50%.

Table 3A: The Gambian dataset (19F)

$\boldsymbol{\nu = 100\%}$	**Mean**	**90%CI**		**Mean**	**90%CI**
$\lambda_{(0,\text{19F})}$	0.15	$(0.08, 0.25)$	$\lambda_{(\text{19F},0)}$	0.21	$(0.08, 0.55)$
$\lambda_{(0,\text{6B})}$	0.35	$(0.28, 0.43)$	$\lambda_{(\text{6B},0)}$	0.40	$(0.30, 0.50)$
$\lambda_{(0,\text{6A})}$	0.28	$(0.22, 0.33)$	$\lambda_{(\text{6A},0)}$	0.49	$(0.36, 0.63)$
$\lambda_{(0,14)}$	0.21	$(0.16, 0.26)$	$\lambda_{(14,0)}$	0.32	$(0.24, 0.42)$
$\lambda_{(0,\text{23F})}$	0.20	$(0.16, 0.26)$	$\lambda_{(\text{23F},0)}$	0.38	$(0.29, 0.50)$
$\lambda_{(0,\text{19A})}$	0.10	$(0.07, 0.13)$	$\lambda_{(\text{19A},0)}$	0.28	$(0.19, 0.37)$
$\lambda_{(0,\text{15B/C})}$	0.07	$(0.05, 0.09)$	$\lambda_{(\text{15B/C},0)}$	0.21	$(0.14, 0.30)$
$\lambda_{(0,\text{35B})}$	0.04	$(0.03, 0.06)$	$\lambda_{(\text{35B},0)}$	0.25	$(0.17, 0.35)$
$\lambda_{(0,3)}$	0.09	$(0.06, 0.13)$	$\lambda_{(3,0)}$	0.48	$(0.30, 0.73)$
$\lambda_{(0,11)}$	0.04	$(0.03, 0.06)$	$\lambda_{(11,0)}$	0.25	$(0.15, 0.37)$
$\lambda_{(0,\text{23B})}$	0.05	$(0.03, 0.07)$	$\lambda_{(\text{23B},0)}$	0.35	$(0.22, 0.51)$
$\lambda_{(0,\text{9N})}$	0.04	$(0.03, 0.06)$	$\lambda_{(\text{9N},0)}$	0.31	$(0.19, 0.46)$
$\lambda_{(0,\text{16F})}$	0.05	$(0.03, 0.07)$	$\lambda_{(\text{16F},0)}$	0.44	$(0.27, 0.64)$
$\lambda_{(0,\text{The rest})}$	0.73	$(0.64, 0.84)$	$\lambda_{(\text{The rest},0)}$	0.56	$(0.47, 0.65)$
$\theta_{(x),(x,\text{19F})}$	0.58	$(0.23, 1.00)$	$\phi_{(\text{19F},x),(x)}$	11.58	$(1.29, 30.72)$
$\theta_{(\text{19F}),(\text{19F},x)}$	0.60	$(0.23, 1.05)$	$\phi_{(x,\text{19F}),(\text{19F})}$	10.73	$(5.01, 19.36)$
$\theta_{(x),(x,y)}$	0.49	$(0.39, 0.59)$	$\phi_{(x,y),(x)}$	13.62	$(9.81, 18.00)$
$\boldsymbol{\nu = 50\%}$	**Mean**	**90%CI**		**Mean**	**90%CI**
$\lambda_{(0,\text{19F})}$	0.17	$(0.11, 0.24)$	$\lambda_{(\text{19F},0)}$	0.24	$(0.13, 0.42)$
$\lambda_{(0,\text{6B})}$	0.31	$(0.25, 0.38)$	$\lambda_{(\text{6B},0)}$	0.40	$(0.31, 0.52)$
$\lambda_{(0,\text{6A})}$	0.23	$(0.19, 0.28)$	$\lambda_{(\text{6A},0)}$	0.46	$(0.36, 0.58)$
$\lambda_{(0,14)}$	0.20	$(0.16, 0.24)$	$\lambda_{(14,0)}$	0.36	$(0.28, 0.44)$
$\lambda_{(0,\text{23F})}$	0.19	$(0.15, 0.24)$	$\lambda_{(\text{23F},0)}$	0.43	$(0.34, 0.54)$
$\lambda_{(0,\text{19A})}$	0.08	$(0.06, 0.10)$	$\lambda_{(\text{19A},0)}$	0.25	$(0.18, 0.35)$
$\lambda_{(0,\text{15B/C})}$	0.08	$(0.06, 0.10)$	$\lambda_{(\text{15B/C},0)}$	0.32	$(0.22, 0.45)$
$\lambda_{(0,\text{35B})}$	0.05	$(0.03, 0.07)$	$\lambda_{(\text{35B},0)}$	0.33	$(0.23, 0.46)$
$\lambda_{(0,3)}$	0.10	$(0.07, 0.14)$	$\lambda_{(3,0)}$	0.75	$(0.51, 1.06)$
$\lambda_{(0,11)}$	0.05	$(0.03, 0.07)$	$\lambda_{(11,0)}$	0.36	$(0.23, 0.51)$
$\lambda_{(0,\text{23B})}$	0.05	$(0.04, 0.07)$	$\lambda_{(\text{23B},0)}$	0.43	$(0.29, 0.60)$
$\lambda_{(0,\text{9N})}$	0.03	$(0.02, 0.05)$	$\lambda_{(\text{9N},0)}$	0.30	$(0.20, 0.43)$
$\lambda_{(0,\text{16F})}$	0.03	$(0.02, 0.04)$	$\lambda_{(\text{16F},0)}$	0.24	$(0.14, 0.36)$
$\lambda_{(0,\text{The rest})}$	0.68	$(0.59, 0.78)$	$\lambda_{(\text{The rest},0)}$	0.59	$(0.50, 0.66)$
$\theta_{(x),(x,\text{19F})}$	0.42	$(0.22, 0.70)$	$\phi_{(\text{19F},x),(x)}$	2.33	$(0.81, 5.00)$
$\theta_{(\text{19F}),(\text{19F},x)}$	0.51	$(0.35, 0.71)$	$\phi_{(x,\text{19F}),(\text{19F})}$	2.33	$(1.56, 3.26)$
$\theta_{(x),(x,y)}$	0.39	$(0.32, 0.46)$	$\phi_{(x,y),(x)}$	3.16	$(2.52, 3.99)$

Table 3B: The Gambian dataset (6B)

$\boldsymbol{\nu} = \mathbf{100\%}$	**Mean**	**90%CI**		**Mean**	**90%CI**
$\lambda_{(0,\mathrm{6B})}$	0.23	$(0.16, 0.30)$	$\lambda_{(\mathrm{6B},0)}$	0.19	$(0.10, 0.31)$
$\lambda_{(0,\mathrm{19F})}$	0.20	$(0.16, 0.24)$	$\lambda_{(\mathrm{19F},0)}$	0.26	$(0.21, 0.32)$
$\lambda_{(0,\mathrm{6A})}$	0.28	$(0.23, 0.32)$	$\lambda_{(\mathrm{6A},0)}$	0.50	$(0.38, 0.64)$
$\lambda_{(0,14)}$	0.21	$(0.18, 0.25)$	$\lambda_{(14,0)}$	0.37	$(0.29, 0.46)$
$\lambda_{(0,\mathrm{23F})}$	0.20	$(0.17, 0.24)$	$\lambda_{(\mathrm{23F},0)}$	0.42	$(0.32, 0.53)$
$\lambda_{(0,\mathrm{19A})}$	0.10	$(0.08, 0.12)$	$\lambda_{(\mathrm{19A},0)}$	0.31	$(0.23, 0.39)$
$\lambda_{(0,\mathrm{15B/C})}$	0.07	$(0.05, 0.09)$	$\lambda_{(\mathrm{15B/C},0)}$	0.25	$(0.17, 0.34)$
$\lambda_{(0,\mathrm{35B})}$	0.05	$(0.03, 0.06)$	$\lambda_{(\mathrm{35B},0)}$	0.30	$(0.21, 0.42)$
$\lambda_{(0,3)}$	0.10	$(0.07, 0.13)$	$\lambda_{(3,0)}$	0.57	$(0.39, 0.83)$
$\lambda_{(0,11)}$	0.05	$(0.03, 0.06)$	$\lambda_{(11,0)}$	0.31	$(0.20, 0.43)$
$\lambda_{(0,\mathrm{23B})}$	0.05	$(0.04, 0.07)$	$\lambda_{(\mathrm{23B},0)}$	0.40	$(0.27, 0.55)$
$\lambda_{(0,\mathrm{9N})}$	0.04	$(0.03, 0.06)$	$\lambda_{(\mathrm{9N},0)}$	0.35	$(0.23, 0.50)$
$\lambda_{(0,\mathrm{16F})}$	0.05	$(0.03, 0.07)$	$\lambda_{(\mathrm{16F},0)}$	0.52	$(0.34, 0.78)$
$\lambda_{(0,\text{The rest})}$	0.71	$(0.62, 0.82)$	$\lambda_{(\text{The rest},0)}$	0.55	$(0.47, 0.64)$
$\theta_{(x),(x,\mathrm{6B})}$	0.84	$(0.57, 1.24)$	$\phi_{(x,\mathrm{6B}),(x)}$	53.69	$(25.41, 117.82)$
$\theta_{(\mathrm{6B}),(x,\mathrm{6B})}$	0.72	$(0.51, 0.97)$	$\phi_{(x,\mathrm{6B}),(\mathrm{6B})}$	34.20	$(20.87, 48.22)$
$\theta_{(x),(x,y)}$	0.34	$(0.27, 0.42)$	$\phi_{(x,y),(x)}$	7.28	$(5.18, 10.37)$
$\boldsymbol{\nu} = \mathbf{50\%}$	**Mean**	**90%CI**		**Mean**	**90%CI**
$\lambda_{(0,\mathrm{6B})}$	0.25	$(0.19, 0.34)$	$\lambda_{(\mathrm{6B},0)}$	0.18	$(0.09, 0.31)$
$\lambda_{(0,\mathrm{19F})}$	0.17	$(0.13, 0.21)$	$\lambda_{(\mathrm{19F},0)}$	0.24	$(0.19, 0.32)$
$\lambda_{(0,\mathrm{6A})}$	0.22	$(0.18, 0.27)$	$\lambda_{(\mathrm{6A},0)}$	0.45	$(0.34, 0.58)$
$\lambda_{(0,14)}$	0.18	$(0.15, 0.22)$	$\lambda_{(14,0)}$	0.36	$(0.27, 0.46)$
$\lambda_{(0,\mathrm{23F})}$	0.18	$(0.14, 0.23)$	$\lambda_{(\mathrm{23F},0)}$	0.44	$(0.35, 0.55)$
$\lambda_{(0,\mathrm{19A})}$	0.08	$(0.06, 0.10)$	$\lambda_{(\mathrm{19A},0)}$	0.26	$(0.19, 0.34)$
$\lambda_{(0,\mathrm{15B/C})}$	0.07	$(0.05, 0.10)$	$\lambda_{(\mathrm{15B/C},0)}$	0.31	$(0.22, 0.43)$
$\lambda_{(0,\mathrm{35B})}$	0.05	$(0.03, 0.06)$	$\lambda_{(\mathrm{35B},0)}$	0.35	$(0.24, 0.47)$
$\lambda_{(0,3)}$	0.10	$(0.07, 0.13)$	$\lambda_{(3,0)}$	0.77	$(0.53, 1.05)$
$\lambda_{(0,11)}$	0.05	$(0.03, 0.07)$	$\lambda_{(11,0)}$	0.38	$(0.25, 0.53)$
$\lambda_{(0,\mathrm{23B})}$	0.05	$(0.03, 0.07)$	$\lambda_{(\mathrm{23B},0)}$	0.45	$(0.31, 0.61)$
$\lambda_{(0,\mathrm{9N})}$	0.03	$(0.02, 0.04)$	$\lambda_{(\mathrm{9N},0)}$	0.31	$(0.19, 0.43)$
$\lambda_{(0,\mathrm{16F})}$	0.02	$(0.01, 0.04)$	$\lambda_{(\mathrm{16F},0)}$	0.25	$(0.15, 0.38)$
$\lambda_{(0,\text{The rest})}$	0.68	$(0.59, 0.79)$	$\lambda_{(\text{The rest},0)}$	0.61	$(0.53, 0.70)$
$\theta_{(x),(x,\mathrm{6B})}$	0.69	$(0.38, 1.07)$	$\phi_{(x,\mathrm{6B}),(x)}$	18.97	$(7.97, 34.08)$
$\theta_{(\mathrm{6B}),(x,\mathrm{6B})}$	0.54	$(0.38, 0.77)$	$\phi_{(x,\mathrm{6B}),(\mathrm{6B})}$	7.12	$(4.26, 10.60)$
$\theta_{(x),(x,y)}$	0.39	$(0.32, 0.47)$	$\phi_{(x,y),(x)}$	2.40	$(1.87, 3.07)$

Table 3. Estimates of acquisition ($\lambda_{(0,.)}$), clearance ($\lambda_{(.,0)}$) and competition parameters ($\theta_{(.,.)}, \phi_{(.,.)}$). The posterior mean and 90% credible intervals (CI; in parenthesis) with the detection sensitivity (ν) of double colonization assumed either 100% or 50%. The target serotype (s) is either 19F in panel A and 6B in panel B.

Competition with very low assumed level of the detection sensitivity. We performed analyses with very low assumed levels of the detection sensitivity (ν), but all model parameters could not be estimated well. We therefore repeated these analyses using a model with a reduced number of parameters. Specifically, we assumed that all serotypes had the same clearance rate and the same competition parameters, i.e., no serotype-specific competition was allowed in the model. The sensitivity parameter ν was assumed to be 15%. The estimated acquisition rates were similar to what were estimated assuming $\nu = 50\%$. The estimated clearance rates were close to what were estimated for the common serotypes in each dataset with $\nu = 50\%$. The estimated levels of competition are presented in Table 4 below.

Table 4: Competition under a reduced model

The Danish dataset					
$\boldsymbol{\nu = 15\%}$	**Mean**	**90%CI**		**Mean**	**90%CI**
$\theta_{(x),(x,y)}$	0.35	$(0.31, 0.39)$	$\phi_{(x,y),(x)}$	0.21	$(0.12, 0.34)$
The American Indian dataset					
$\boldsymbol{\nu = 15\%}$	**Mean**	**90%CI**		**Mean**	**90%CI**
$\theta_{(x),(x,y)}$	0.62	$(0.41, 0.86)$	$\phi_{(x,y),(x)}$	0.74	$(0.49, 1.06)$
The Gambian dataset					
$\boldsymbol{\nu = 15\%}$	**Mean**	**90%CI**		**Mean**	**90%CI**
$\theta_{(x),(x,y)}$	0.55	$(0.47, 0.62)$	$\phi_{(x,y),(x)}$	1.78	$(1.49, 2.01)$

Table 4. Estimates of competition parameters ($\theta_{(x,y)}, \phi_{(x,y)}$) under a reduced model in which all serotypes share the same competition parameters and clearance rates. The posterior mean and 90% credible intervals (CI; in parenthesis) with the detection sensitivity (ν) of double colonization assumed 15%.

We also compared the model predictions with $\nu = 15\%$ to the observed data. Briefly, the model does not allow competition in acquisition to be close to 1 because the turnover of serotypes would then be faster than what was observed.

3 eAppendix C: Model assessment

Figures 1, 2 and 3 present the observed transition probabilities between the different colonization states for the most common sampling intervals in the Danish, American Indian, and the Gambian datasets, respectively. In addition, the posterior predictive transition probabilities are presented, based on the model with the either perfect ($\nu = 100\%$) or imperfect sensitivity ($\nu = 50\%$) to detect double colonization assumed. For $\nu = 50\%$, the transition probabilities which take into account the detection sensitivity, i.e., not the true underlying ones, are presented. Transition probabilities are determined between the following five aggregated states: non-colonized (state 0), the target serotype, any non-target type, doubly colonized states (target and non-target; two non-target types). The target type is not present with the American Indian data because target-specific rates could not be identified.

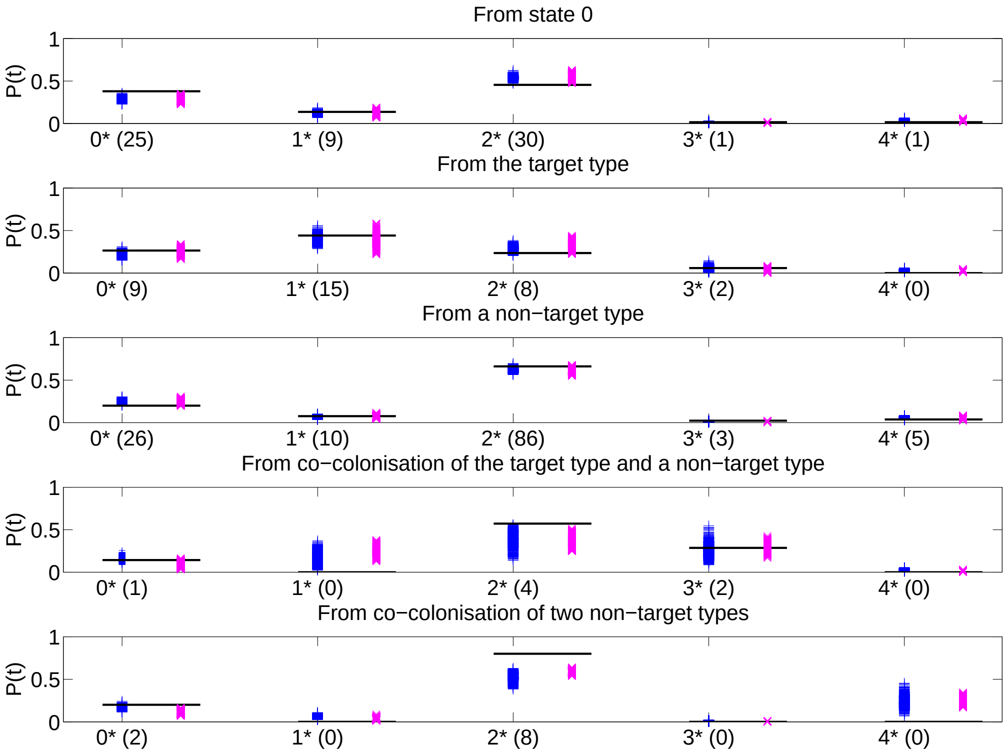

Figure 1: The observed transition probabilities between the different colonization states (P(t)) for the most common time interval (30 days) in the Danish dataset and serotype 23F as the target type (horizontal line). The posterior predictive distributions of the transition probabilities are indicated with blue '+' (the model with 100% sensitivity of detection) or magenta 'x' (50% sensitivity). 0^* : To state 0. 1^* : To the target type. 2^* : To a non-target type. 3^* : To co-colonization of the target type and a non-target type. 4^* : To co-colonization of two non-target types.

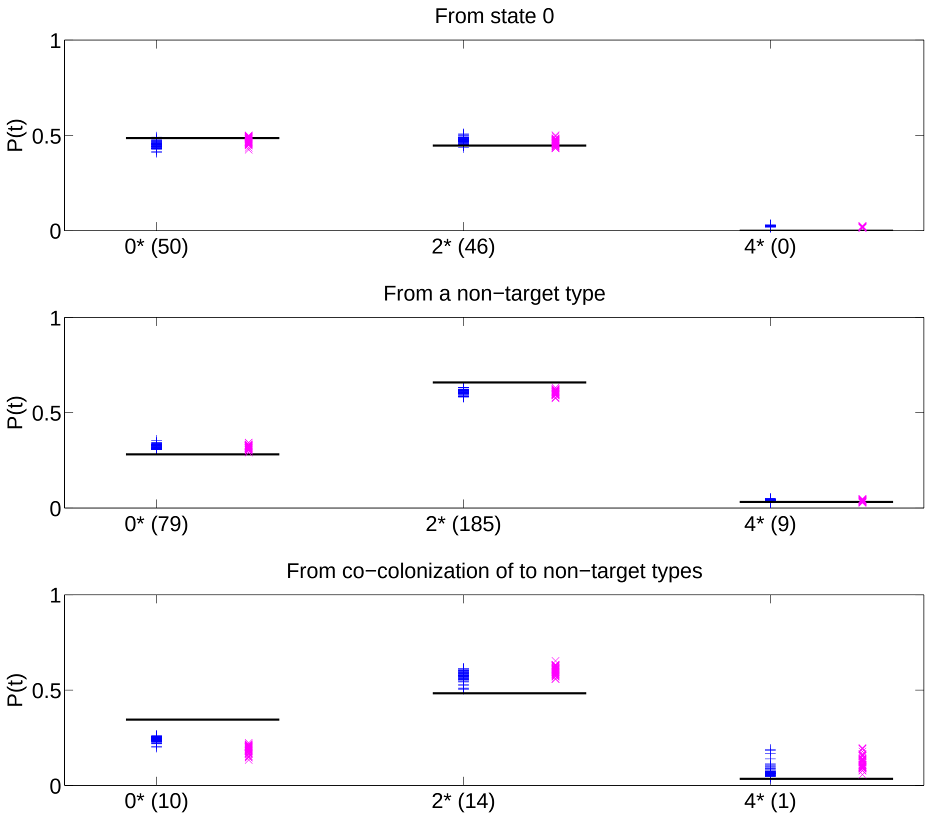

Figure 2: The observed transition probabilities between the different colonization states (P(t)) for the most common time interval (35 days) in the American Indian dataset (horizontal line). The posterior predictive distributions of the transition probabilities are indicated with blue '+' (the model with 100% sensitivity of detection) or magenta 'x' (50% sensitivity). 0^* : To state 0. 2^* : To a non-target type. 4^* : To co-colonization of two non-target types

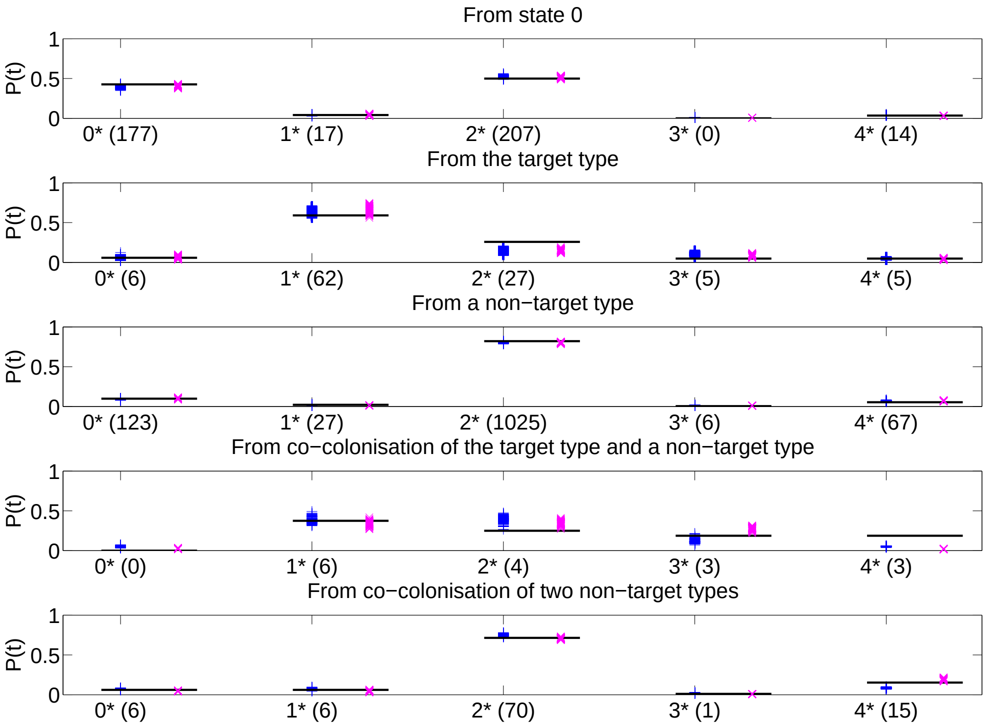

Figure 3: The observed transition probabilities between the different colonization states (P(t)) for the most common time interval (14 days) in the Gambian dataset and serotype 19F as the target type (horizontal line). The posterior predictive distributions of the transition probabilities are indicated with blue '+' (the model with 100% sensitivity of detection) or magenta 'x' (50% sensitivity). (50% sensitivity 0^* : To state 0. 1^* : To the target type. 2^* : To a non-target type. 3^* : To co-colonization of the target type and a non-target type. 4^* : To co-colonization of two non-target types

Calculation of the prevalence of colonization and the proportion of doubly colonized samples in the absence of competition. Assuming that the total prevalence and the serotype distribution are approximately the observed ones, the proportion of multiple colonization among the positive samples under the assumption of no competition can be calculated. Denote $p_i = N_i/N$, where N_i is the number of isolates of serotype i and N the total number of samples in the dataset, and $P_1 = \sum_i p_i$, $P_2 = \sum_i \sum_{j>i} p_i p_j$, and $P_3 = \sum_i \sum_{j>i} \sum_{k>j} p_i p_j p_k$, ..., where for each dataset we have $P_{i:i>3} \approx 0$. Thus, by the additive law of probability for independent events, the total prevalence of colonization is approximately $P_1 - P_2 + P_3$ and the prevalence of multiple colonization $P_2 - P_3$, and the proportion of multiple colonization of total colonization $(P_2 - P_3)/(P_1 - P_2 + P_3)$.

By scaling the observed p_i in each dataset so that the total prevalence $(P_1 - P_2 + P_3)$ is the observed one, we obtain the theoretical proportion of multiple colonization of these in the absence of competition. For the Danish dataset this is $(P_2 - P_3)/(P_1 - P_2 + P_3) \approx 55\%$, for the American Indian dataset $(P_2 - P_3)/(P_1 - P_2 + P_3) \approx 40\%$, and for the Gambian dataset $(P_2 - P_3)/(P_1 - P_2 + P_3) \approx 62\%$.

Assessment of exposure as a confounder. The estimation of competition in acquisition could be confounded, if some unadjusted background variables were associated with both the colonization status (colonized/non-colonized) and exposure to acquisition. For instance, individuals could belong to subpopulations clearly dominated by different serotypes (micro-epidemics). We employed the Danish data to investigate whether such associations could be found. In that dataset, repeated observations came from three different day-care centers. Earlier studies have indicated the key role of day care centers in generating micro-epidemics.[4]

We calculated first the proportion of non-colonized samples that were collected with and without serotype 23F (the target serotype in the analysis) present at the same time at the same day care center. Second, we calculated the proportion of samples found to be singly colonized with other types than 23F, again with and without 23F being present at the same time at the same day care center. The results are shown in Table 5 below.

	$\mathbf{N}_0$	$\mathbf{N}_y$	**Total**
Exposure = 0	73(73/165 = 44%)	110(110/340 = 32%)	183
Exposure = 1	92(92/165 = 56%)	230(230/340 = 68)%	322
Total	165	340	505

N_0:Number of observations of the non-colonized state 0.

N_y:Number of observations of singly colonized states other than 23F.

Exposure: exposure to serotype 23F (23F present at the same DCC at the same sampling round, 1=yes, 0=no).

Table 5. The number of observations of different states in the Danish dataset stratified by the presence of serotype 23F (yes/no at the respective sampling round in the same DCC).

The proportion of samples found non-colonized and without exposure to 23F was 44% of all non-colonized samples. The corresponding proportion of singly colonized samples (other than 23F) was 32%. If colonization with other types than 23F was associated with 23F not being present at the same time, the latter proportion should be larger. Since this was not the case, this means that those colonized with other than the target serotype were actually more often exposed to 23F than the non-colonized. This is against the hypothesis of confounding bias due to micro-epidemics. The reason why other types were present at the same time as 23F could be e.g. seasonal effects. Perhaps more likely, it is just random variation as for all serotypes the average difference of the corresponding proportions was only 4 percentage points.

References

[1] Nagelkerke NJ, Chunge RN, Kinoti SN. Estimation of parasitic infection dynamics when detectability is imperfect. *Statistics in Medicine* 1990; **9**(10): 1211-1219.

[2] Cappé O, Moulines E, Rydn T. Inference in Hidden Markov Models. *Springer* 2005.

[3] Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**: 457-511.

[4] Hoti F, Erästö P, Leino T, et al. Outbreaks of Streptococcus pneumoniae carriage in day care cohorts in Finland - implications for elimination of transmission. *BMC Infect Dis.* 2009; **9**(102).