eAppendix to Lipsitch et al. Estimating rates of carriage acquisition and clearance and competitive ability for pneumococcal serotypes in Kenya with a Markov transition model.

eTable 1: Observations used in the analysis

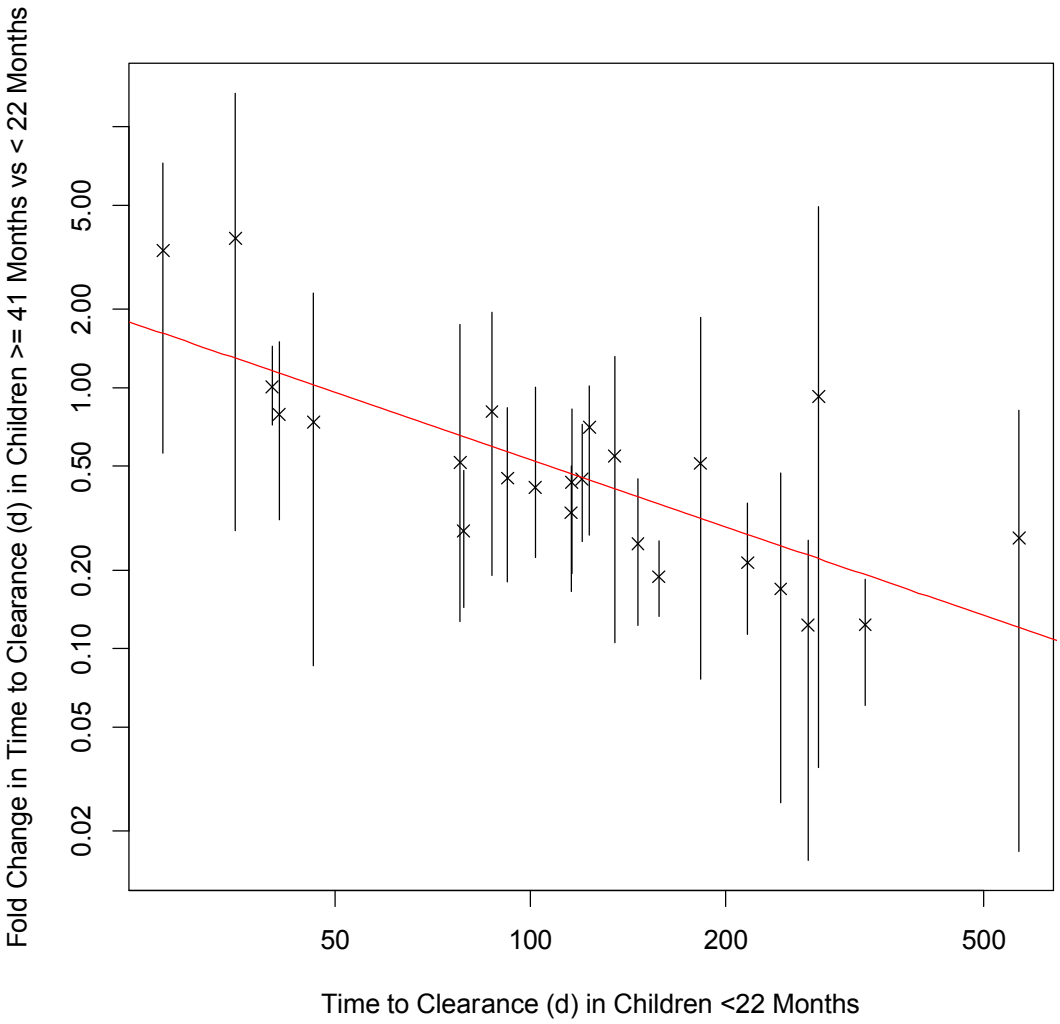
Observation	n (total = 2840 persons)
Uncolonized at baseline	972 (34% of all persons)
Colonized at baseline, <3 swabs (could not evaluate for time to event)	390 (14% all persons)
Colonized at baseline, 3 or more swabs (could be evaluated for time to event)	1478 (52% of all persons)
Colonized at baseline, no transition observed before end of follow up	421 (28% of persons with 3 or more swabs)
Clearance observed	444 (30% of persons with 3 or more swabs) Of these: 292 (66%) with two consecutive negative swabs at the end of carriage
Change of serotype observed	613 (42% of persons with 3 or more swabs) Of these: 322 (53%) with two consecutive identical swabs showing same new serotype

eTable 2: Transitions between pairs of serotypes that occurred more often than expected by chance in 95% of cases.

Serotype Transition	Expectation	Observation minus expectation
15B -> 15C	**1.765329296**	**16.2346707**
15C -> 15B	**1.635124905**	**10.36487509**
19F -> 35B	7.249810749	9.750189251
35B -> 35A	**0.698713096**	**9.301286904**
6B -> 6A	**10.78046934**	**9.219530659**
23F -> 6A	8.965934898	9.034065102
6A -> 6B	**9.286146858**	**7.713853142**
23A -> 23B	**0.882664648**	**7.117335352**
7C -> 20	0.508705526	6.491294474
23F -> 23A	**1.398940197**	**5.601059803**
35A -> 35B	**0.554125662**	**5.445874338**
6B -> 14	3.822861469	5.177138531
7C -> 19F	3.342922029	4.657077971
23B -> 23A	**0.749432248**	**4.250567752**
19A -> 19B	**0.849356548**	**4.150643452**
35B -> 9V	1.881150643	4.118849357
20 -> 7C	0.472369417	3.527630583
10A -> 6A	4.482967449	3.517032551
19F -> 18C	2.020439061	2.979560939
19F -> 33D	2.020439061	2.979560939
23F -> 21	1.144587434	2.855412566
33B -> 33D	0.167297502	2.832702498
15A -> 10A	1.308099924	2.691900076
9V -> 15C	1.323996972	2.676003028
35F -> 35B	**0.369417108**	**2.630582892**
35B -> 24F	0.42997729	2.57002271
16F -> 6B	1.448902347	2.551097653
23B -> 23F	**2.452687358**	**2.547312642**
34 -> 11A	1.53671461	2.46328539
21 -> 14	0.567751703	2.432248297
18F -> 15A	0.046934141	1.953065859
33D -> 33B	0.063588191	1.936411809
10A -> 10F	0.127176382	1.872823618
18C -> 24F	0.139288418	1.860711582
20 -> 4	0.163512491	1.836487509
35B -> 35F	**0.268735806**	**1.731264194**
6B -> 5	0.535200606	1.464799394
28A -> 28F	0.002271007	0.997728993
19C -> 19A	**0.070401211**	**0.929598789**

Bold rows indicate within-serogroup pairs that may reflect high-frequency switching or errors in the last step of serotyping.

eFigure 1: Reduction in time to clearance from the youngest to the oldest terciles, versus time to clearance in the youngest tercile (log-log plot). The strong negative correlation suggests that with age, the more rapid immune clearance of carriage has a greater proportional effect on long-carried serotypes.



Electronic availability of data:

We have included R objects that correspond to each of the model fits mentioned in this paper.

The full-data fit is called multi_res_fit_sigma_revision.RData , and the separate fits by age tercile have similar names with the terciles of age (1=youngest; 3=oldest).

When loaded into R, these appear as an object named "fit". The "fit" object has several slots:

fit$res: Table of parameter estimates with asymptotic 95% confidence intervals for each.
fit$init: Vector of fitted prevalence estimates for each of the serotypes and the uncolonized state.
fit$sigma: Asymptotic covariance matrix which can be used to simulate from the asymptotic sampling distribution of the maximum likelihood log-parameter estimators (or the approximate posterior distribution of the parameters from a Bayesian perspective) to compute the sampling distribution of other quantities of interest (details below). Alternatively, this covariance matrix can be used for Delta method calculations.
fit$loglik: Log-likelihood value at the parameter estimates.
fit$optim.res: List returned by the optim routine used to fit the model. Contains additional convergence information and the numerical Hessian.
fit$init: Initial value of each of the parameters when the optimizer was initialized.

Example:
To simulate from the asymptotic sampling distribution of the whole set of parameters:

```
> library(mvtnorm)
> rmvnorm(100, log(fit$res[1,]), fit$sigma))
```

Note that the covariance matrix corresponds to the sampling distribution of the log-parameters, so the normal distributions is centered at the log of the maximum likelihood estimates.