Statistical Analysis Plan for SCOPE

Dan Weinberger

2022-09-22

## Overview

The SCOPE study enrolled adults 60 years of age and older in greater New Haven over the course of 2 years (October 2020-September 2022). Individuals were followed longitudinally over a 10 week period (6 visits, 2 weeks apart). At each study visit, individuals provided a saliva sample, and at the final visit, they provided a urine sample. They also completed a questionnaire about their activities and contacts with others. A unique aspect of this study is that subjects were enrolled as pairs living in the same household, and these two people were the only members of the household.

## Study objectives

The goals of the study are to quantify the acquisition and clearance rates of pneumococcus among older adults living in a community setting, to identify risk factors for colonization, and to quantify within-household transmission of pneumococcus.

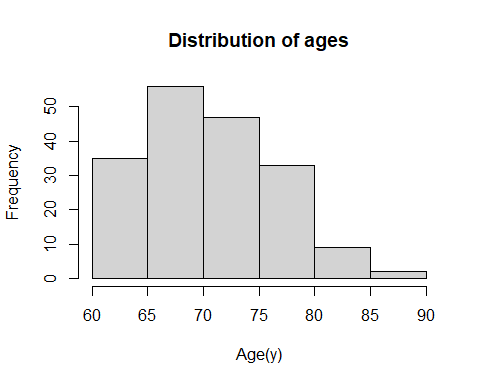
## Laboratory methods

In the laboratory, saliva was culture enriched and tested for *lytA* and *piaB with qPCR.* A Ct value of <40 was considered to be positive. Because piaB is a highly specific marker for pneumococcus and because piaB was more sensitive than the lytA assay at low concentration, we considered those who were positive for piaB to be positive for pneumococcus. To reduce the possibility of false positivity, any samples in the range of 35-40 Ct were retested to confirm the positive. In the case of discrepant results, they were tested a 3rd time as a tie breaker.

## Analysis plan

### Descriptive statistics

Age summary

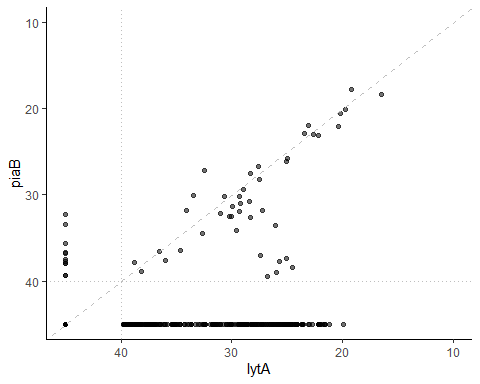


## # A tibble: 1 x 3  
## mean min max  
## <dbl> <dbl> <dbl>  
## 1 71.0 60 86

|  | Overall |
| --- | --- |
|  | (N=182) |
| Age |  |
| Mean (SD) | 70.9 (5.89) |
| Median [Min, Max] | 70.0 [60.0, 86.0] |
| Missing | 1 (0.5%) |
| Gender |  |
| F | 93 (51.1%) |
| M | 89 (48.9%) |
| Race |  |
| asian | 4 (2.2%) |
| black or african american | 6 (3.3%) |
| unknown | 13 (7.1%) |
| white | 158 (86.8%) |
| Missing | 1 (0.5%) |
| Ethnicity |  |
| hispanic or latino | 1 (0.5%) |
| non-hispanic | 173 (95.1%) |
| unknown | 7 (3.8%) |
| Missing | 1 (0.5%) |
| activity\_family |  |
| Mean (SD) | 5220 (5010) |
| Median [Min, Max] | 10000 [0, 10000] |
| Missing | 2 (1.1%) |
| activity\_friends |  |
| Mean (SD) | 5220 (5010) |
| Median [Min, Max] | 10000 [0, 10000] |
| Missing | 2 (1.1%) |
| activity\_community\_center |  |
| Mean (SD) | 5220 (5010) |
| Median [Min, Max] | 10000 [0, 10000] |
| Missing | 2 (1.1%) |

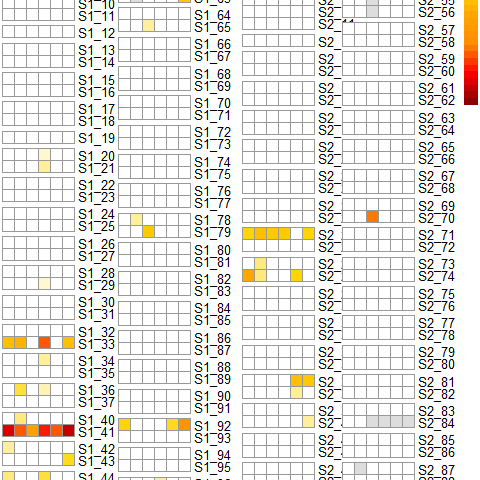
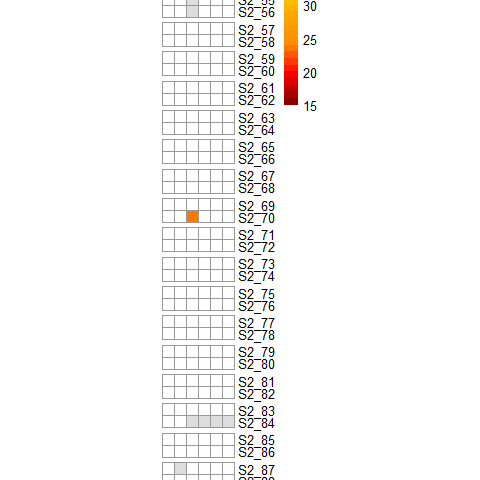
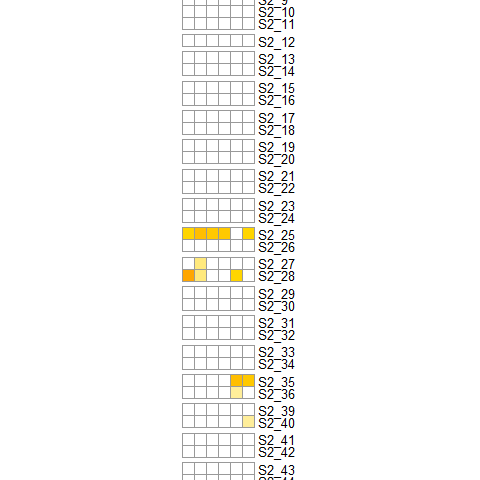
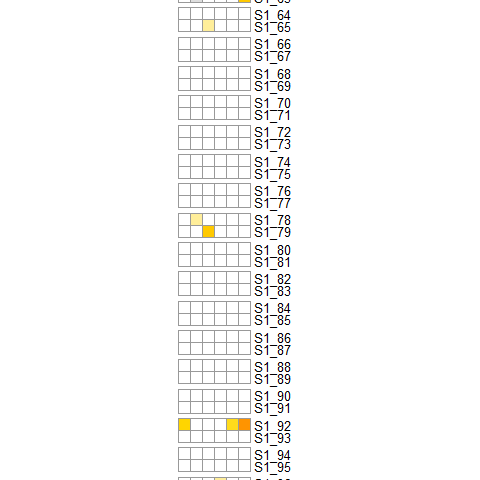
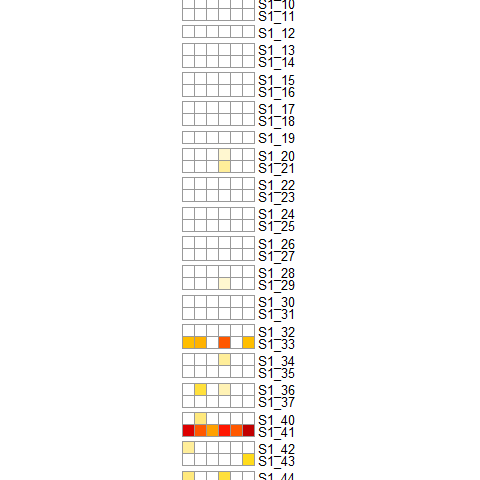
* lytA vs piaB Ct values

## Warning: Removed 2 rows containing missing values (geom\_point).

 - Heatmap of positivity for piaB

## Warning in heat.fun(target = "piab"): NAs introduced by coercion

## Warning: `group\_indices()` was deprecated in dplyr 1.0.0.  
## Please use `cur\_group\_id()` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was generated.



* Percent of tests that were positive, overall

| N\_piaB | N\_test | Percent piab |
| --- | --- | --- |
| 54 | 1,088 | 4.96 |

- and by year

| season | N\_piaB | N\_test | Pct\_piab |
| --- | --- | --- | --- |
| S1 | 37 | 567 | 6.5 |
| S2 | 17 | 521 | 3.3 |

* how many households? How many people?

## # A tibble: 1 x 2  
## N\_hh N\_people  
## <int> <int>  
## 1 93 184

* Percent of people who were positive (period prevalence)

overall

| N\_people | N\_pos | Percent Positive People |
| --- | --- | --- |
| 184 | 30 | 16.3 |

| season | N\_people | N\_pos | Percent Positive People |
| --- | --- | --- | --- |
| S1 | 95 | 20 | 21.1 |
| S2 | 89 | 10 | 11.2 |

* Percent of households that were positive

| season | N\_households | N\_pos | Percent Positive HH |
| --- | --- | --- | --- |
| S1 | 48 | 15 | 31.2 |
| S2 | 45 | 8 | 17.8 |

* Percent positive and average Ct, by sex

| Gender | season | N\_piaB | N\_test | Pct\_piab |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 | 18 | 294 | 6.1 |
|  | S2 | 14 | 268 | 5.2 |
| M |  |  |  |  |
|  | S1 | 19 | 277 | 6.9 |
|  | S2 | 3 | 252 | 1.2 |

or period prevalence by sex and season

| Gender | season | N\_pos\_people | N\_people | period\_prev |
| --- | --- | --- | --- | --- |
| F | S1 | 10 | 49 | 20.4 |
| F | S2 | 7 | 45 | 15.6 |
| M | S1 | 10 | 46 | 21.7 |
| M | S2 | 3 | 43 | 7.0 |

* Percent positive by symptoms

| recent\_nasal | N\_piaB | N\_test | Pct\_piab |
| --- | --- | --- | --- |
| No |  |  |  |
|  | 46 | 881 | 5.2 |
| Yes |  |  |  |
|  | 4 | 110 | 3.6 |
| Missing |  |  |  |
|  | 4 | 100 | 4.0 |

| recent\_runny\_nose | N\_piaB | N\_test | Pct\_piab |
| --- | --- | --- | --- |
| No |  |  |  |
|  | 45 | 892 | 5.0 |
| Yes |  |  |  |
|  | 5 | 99 | 5.1 |
| Missing |  |  |  |
|  | 4 | 100 | 4.0 |

Recent Abx

| recent\_abx | N\_piaB | N\_test | Pct\_piab |
| --- | --- | --- | --- |
| No |  |  |  |
|  | 45 | 850 | 5.3 |
| Yes |  |  |  |
|  | 2 | 67 | 3.0 |
| Missing |  |  |  |
|  | 7 | 174 | 4.0 |

an analysis of positivity by vax status? PPV status? PCV status? PPV/PCV status? Flu+PPV/PCV? Note season 2 questionaire only asked about **recent** vaccine receipt (in past 2 weeks); does not ask about whether ever vaccinated. we have permission to contact PCPs to obtain vax info but have not done this.

| recent\_pneumo\_vax | N\_piaB | N\_test | Pct\_piab |
| --- | --- | --- | --- |
| 0 |  |  |  |
|  | 14 | 405 | 3.5 |
| 1 |  |  |  |
|  | 0 | 1 | 0.0 |
| Missing |  |  |  |
|  | 40 | 685 | 5.8 |

| recent\_flu\_vax | N\_piaB | N\_test | Pct\_piab |
| --- | --- | --- | --- |
| 0 |  |  |  |
|  | 14 | 396 | 3.5 |
| 1 |  |  |  |
|  | 0 | 10 | 0.0 |
| Missing |  |  |  |
|  | 40 | 685 | 5.8 |

* Percent positive by calendar month
* Percent of individuals with contacts, by age of contacts

## Activities by year

Activity with family?

| Gender | season | N\_Activity\_family | N | Pct |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 | 38 | 216 | 17.59 |
|  | S2 | 133 | 252 | 52.78 |
| M |  |  |  |  |
|  | S1 | 33 | 200 | 16.50 |
|  | S2 | 118 | 236 | 50.00 |

contact <5 years

| Gender | season | N\_contact\_u5 | N | Pct |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 | 37 | 233 | 15.88 |
|  | S2 | 55 | 252 | 21.83 |
| M |  |  |  |  |
|  | S1 | 26 | 206 | 12.62 |
|  | S2 | 49 | 236 | 20.76 |

Child contact <12 m

| Gender | season | N\_contact\_u12m | N | Pct |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 | 22 | 233 | 9.4 |
|  | S2 | 26 | 252 | 10.3 |
| M |  |  |  |  |
|  | S1 | 10 | 206 | 4.9 |
|  | S2 | 17 | 236 | 7.2 |

Child contact 12-23m

| Gender | season | N\_contact\_12\_23m | N | Pct |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 | 5 | 233 | 2.15 |
|  | S2 | 28 | 252 | 11.11 |
| M |  |  |  |  |
|  | S1 | 3 | 206 | 1.46 |
|  | S2 | 29 | 236 | 12.29 |

Child contact 24-59 m

| Gender | season | N\_contact\_24\_59m | N | Pct |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 | 17 | 233 | 7.30 |
|  | S2 | 29 | 252 | 11.51 |
| M |  |  |  |  |
|  | S1 | 16 | 206 | 7.77 |
|  | S2 | 23 | 236 | 9.75 |

Child contact 5-9y

| Gender | season | N\_contact\_5\_9y | N | Pct |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 | 41 | 233 | 17.60 |
|  | S2 | 50 | 252 | 19.84 |
| M |  |  |  |  |
|  | S1 | 26 | 206 | 12.62 |
|  | S2 | 43 | 236 | 18.22 |

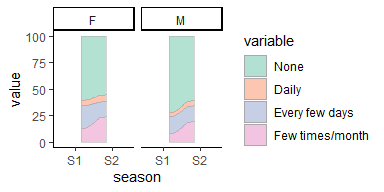
Child contact >10y

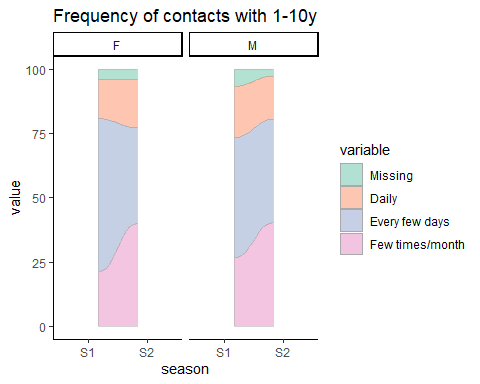
| Gender | season | N\_contact\_10y\_older | N | Pct |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 | 30 | 233 | 12.88 |
|  | S2 | 53 | 252 | 21.03 |
| M |  |  |  |  |
|  | S1 | 20 | 206 | 9.71 |
|  | S2 | 38 | 236 | 16.10 |

How does intensity of child contact vary by season

0= none reported, 1=Daily; 2=every few days; 3= once or twice a month

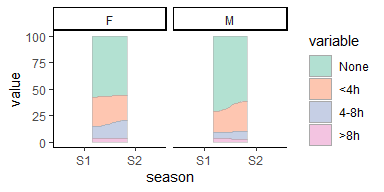
| Gender | season | child\_contact\_often | N | Percent |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 |  |  |  |
|  |  | 0 | 131 | 60.65 |
|  |  | 1 | 11 | 5.09 |
|  |  | 2 | 47 | 21.76 |
|  |  | 3 | 27 | 12.50 |
|  | S2 |  |  |  |
|  |  | 0 | 139 | 55.38 |
|  |  | 1 | 16 | 6.37 |
|  |  | 2 | 36 | 14.34 |
|  |  | 3 | 60 | 23.90 |
| M |  |  |  |  |
|  | S1 |  |  |  |
|  |  | 0 | 140 | 72.16 |
|  |  | 1 | 7 | 3.61 |
|  |  | 2 | 32 | 16.49 |
|  |  | 3 | 15 | 7.73 |
|  | S2 |  |  |  |
|  |  | 0 | 143 | 60.59 |
|  |  | 1 | 12 | 5.08 |
|  |  | 2 | 35 | 14.83 |
|  |  | 3 | 46 | 19.49 |

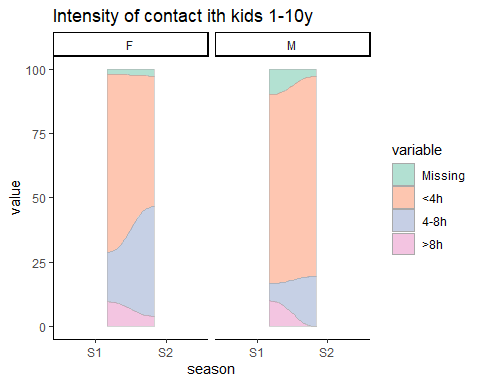


Just among those who have contact with kids 1-10y 

Hours with kids? 1= < 4 hours (just a morning or afternoon or evening) 2= 4-8 hours (full day) 3= 8+ hours (longer visit such as day care or overnight)

| Gender | season | child\_contact\_hours | N | Percent |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 |  |  |  |
|  |  | 0 | 124 | 57.41 |
|  |  | 1 | 61 | 28.24 |
|  |  | 2 | 24 | 11.11 |
|  |  | 3 | 7 | 3.24 |
|  | S2 |  |  |  |
|  |  | 0 | 140 | 55.78 |
|  |  | 1 | 60 | 23.90 |
|  |  | 2 | 42 | 16.73 |
|  |  | 3 | 9 | 3.59 |
| M |  |  |  |  |
|  | S1 |  |  |  |
|  |  | 0 | 138 | 71.13 |
|  |  | 1 | 39 | 20.10 |
|  |  | 2 | 10 | 5.15 |
|  |  | 3 | 7 | 3.61 |
|  | S2 |  |  |  |
|  |  | 0 | 145 | 61.44 |
|  |  | 1 | 67 | 28.39 |
|  |  | 2 | 18 | 7.63 |
|  |  | 3 | 6 | 2.54 |



And just for kids with contact with 1-10 year olds 

Recent runny nose?

| Gender | season | recent\_runny\_nose | N | Percent |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 |  |  |  |
|  |  | 9999 | 30 | 10.20 |
|  |  | No | 253 | 86.05 |
|  |  | Yes | 11 | 3.74 |
|  | S2 |  |  |  |
|  |  | 9999 | 19 | 7.09 |
|  |  | No | 204 | 76.12 |
|  |  | Yes | 45 | 16.79 |
| M |  |  |  |  |
|  | S1 |  |  |  |
|  |  | 9999 | 35 | 12.64 |
|  |  | No | 226 | 81.59 |
|  |  | Yes | 16 | 5.78 |
|  | S2 |  |  |  |
|  |  | 9999 | 16 | 6.35 |
|  |  | No | 209 | 82.94 |
|  |  | Yes | 27 | 10.71 |

| Gender | season | recent\_nasal | N | Percent |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 |  |  |  |
|  |  | 9999 | 30 | 10.20 |
|  |  | No | 244 | 82.99 |
|  |  | Yes | 20 | 6.80 |
|  | S2 |  |  |  |
|  |  | 9999 | 19 | 7.09 |
|  |  | No | 208 | 77.61 |
|  |  | Yes | 41 | 15.30 |
| M |  |  |  |  |
|  | S1 |  |  |  |
|  |  | 9999 | 35 | 12.64 |
|  |  | No | 231 | 83.39 |
|  |  | Yes | 11 | 3.97 |
|  | S2 |  |  |  |
|  |  | 9999 | 16 | 6.35 |
|  |  | No | 198 | 78.57 |
|  |  | Yes | 38 | 15.08 |

Recent antibiotics?

| Gender | season | recent\_abx | N | Percent |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 |  |  |  |
|  |  | 9999 | 31 | 10.54 |
|  |  | No | 252 | 85.71 |
|  |  | Yes | 11 | 3.74 |
|  | S2 |  |  |  |
|  |  | 9999 | 53 | 19.78 |
|  |  | No | 191 | 71.27 |
|  |  | Yes | 24 | 8.96 |
| M |  |  |  |  |
|  | S1 |  |  |  |
|  |  | 9999 | 39 | 14.08 |
|  |  | No | 223 | 80.51 |
|  |  | Yes | 15 | 5.42 |
|  | S2 |  |  |  |
|  |  | 9999 | 51 | 20.24 |
|  |  | No | 184 | 73.02 |
|  |  | Yes | 17 | 6.75 |

The coding on this needs to be fixed.

| Gender | season | recent\_covid\_vax | N | Percent |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 |  |  |  |
|  |  | 9,999 | 294 | 100.00 |
|  | S2 |  |  |  |
|  |  | 0 | 203 | 75.75 |
|  |  | 1 | 6 | 2.24 |
|  |  | 9,999 | 59 | 22.01 |
| M |  |  |  |  |
|  | S1 |  |  |  |
|  |  | 9,999 | 277 | 100.00 |
|  | S2 |  |  |  |
|  |  | 0 | 194 | 76.98 |
|  |  | 1 | 3 | 1.19 |
|  |  | 9,999 | 55 | 21.83 |

## Percent positive by different characteristics?

Are both peoepl in the HH positive? 0= neither positive; 1= 1 person positive, 2=both positive

| season | N\_pos\_time | N |
| --- | --- | --- |
| S1 |  |  |
|  | 0 | 253 |
|  | 1 | 31 |
|  | 2 | 3 |
| S2 |  |  |
|  | 0 | 248 |
|  | 1 | 13 |
|  | 2 | 2 |

## Proportion positive, stratified by contacts

- Contact with children \<5 years

Between seasons, among people without reported child contact, prevalence is much lower; it is similar among those with a reported contact

ANY contact

| child\_contact | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- |
| 0 |  |  |  |
|  | 532 | 12 | 2.3 |
| 1 |  |  |  |
|  | 365 | 31 | 8.5 |
| 9,999 |  |  |  |
|  | 194 | 11 | 5.7 |

| child\_contact\_u5 | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- |
| 0 |  |  |  |
|  | 760 | 28 | 3.7 |
| 1 |  |  |  |
|  | 167 | 19 | 11.4 |
| 9,999 |  |  |  |
|  | 164 | 7 | 4.3 |

- Contact with children \<1 year, 1-2 years, 2-4 years, 5-9 years, 10 years and older  
  
<12m

| child\_contact\_u12m | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- |
| 0 |  |  |  |
|  | 852 | 40 | 4.7 |
| 1 |  |  |  |
|  | 75 | 7 | 9.3 |
| 9,999 |  |  |  |
|  | 164 | 7 | 4.3 |

12-23m

| child\_contact\_13\_23m | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- |
| 0 |  |  |  |
|  | 862 | 41 | 4.8 |
| 1 |  |  |  |
|  | 65 | 6 | 9.2 |
| 9,999 |  |  |  |
|  | 164 | 7 | 4.3 |

| child\_contact\_24\_59m | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- |
| 0 |  |  |  |
|  | 842 | 34 | 4.0 |
| 1 |  |  |  |
|  | 85 | 13 | 15.3 |
| 9,999 |  |  |  |
|  | 164 | 7 | 4.3 |

| child\_contact\_5\_10y | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- |
| 0 |  |  |  |
|  | 767 | 29 | 3.8 |
| 1 |  |  |  |
|  | 160 | 18 | 11.2 |
| 9,999 |  |  |  |
|  | 164 | 7 | 4.3 |

| child\_contact\_over10y | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- |
| 0 |  |  |  |
|  | 786 | 40 | 5.1 |
| 1 |  |  |  |
|  | 141 | 7 | 5.0 |
| 9,999 |  |  |  |
|  | 164 | 7 | 4.3 |

### intensity of contact

Seems to show people with less intense contact have higher prevalence. This seems counter-intuitive, until you look at the next 2 tables, which show that prevalence is higher in people with more frequent contact; and people with more frequent contact tend to have fewer contact hours per visit. This suggest frequency is the driver.

1, < 4 hours (just a morning or afternoon or evening) | 2, 4-8 hours (full day) | 3, 8+ hours (longer visit such as day care or overnight)

| child\_contact\_hours | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- |
| 1 | 252 | 28 | 11.1 |
| 2 | 107 | 7 | 6.5 |
| 3 | 35 | 1 | 2.9 |
| 999 | 697 | 18 | 2.6 |

frequency of contact vs pneumo positivity 1, Daily | 2, Every few days | 3, Once or twice a month This makes sense–more common contatct, more carriage

N\_pos is the number of positives for that level of contact frequency; pct\_pos is percent positive for that level

| child\_contact\_often | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- |
| 1 | 46 | 7 | 15.2 |
| 2 | 153 | 18 | 11.8 |
| 3 | 149 | 6 | 4.0 |
| 999 | 743 | 23 | 3.1 |

What is relationship between frequency of child contact and intensity? Higher frequency; less intensity per visit

##   
## 1 2 3   
## 220 92 28

##   
## 1 2 3   
## 46 151 143

##   
## 1 2 3  
## 1 28 8 10  
## 2 92 53 6  
## 3 100 31 12

Out of curiousity, break down the table of pneumo poisitivity by contact frequency AND gender. This is interesting; women who have daily contact have super high prevalence; men not so much.

| Gender | child\_contact\_often | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | 1 | 27 | 6 | 22.2 |
|  | 2 | 85 | 8 | 9.4 |
|  | 3 | 87 | 4 | 4.6 |
|  | 999 | 363 | 14 | 3.9 |
| M |  |  |  |  |
|  | 1 | 19 | 1 | 5.3 |
|  | 2 | 68 | 10 | 14.7 |
|  | 3 | 62 | 2 | 3.2 |
|  | 999 | 380 | 9 | 2.4 |

| Gender | child\_contact\_hours | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | 1 | 133 | 17 | 12.8 |
|  | 2 | 76 | 3 | 3.9 |
|  | 3 | 19 | 0 | 0.0 |
|  | 999 | 177 | 9 | 5.1 |
|  |  | 157 | 3 | 1.9 |
| M |  |  |  |  |
|  | 1 | 119 | 11 | 9.2 |
|  | 2 | 31 | 4 | 12.9 |
|  | 3 | 16 | 1 | 6.2 |
|  | 999 | 202 | 6 | 3.0 |
|  |  | 161 | 0 | 0.0 |

Lets create a combined index..groups 7 and 8 have highest prevalence; this is people who have daily or every few day contact for short duration

| intensity\_index | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- |
| 1 | 10 | 1 | 10.0 |
| 2 | 6 | 0 | 0.0 |
| 3 | 12 | 0 | 0.0 |
| 4 | 8 | 0 | 0.0 |
| 5 | 53 | 5 | 9.4 |
| 6 | 31 | 2 | 6.5 |
| 7 | 28 | 6 | 21.4 |
| 8 | 92 | 13 | 14.1 |
| 9 | 100 | 3 | 3.0 |
| 999 | 747 | 24 | 3.2 |
|  | 4 | 0 | 0.0 |

Prevalence by sex. Seems that men and women had similar prevalence in Season 1; Women had similar prevalence again in Season 2, men had very little carriage.

| Gender | season | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- | --- |
| F |  |  |  |  |
|  | S1 | 294 | 18 | 6.1 |
|  | S2 | 268 | 14 | 5.2 |
| M |  |  |  |  |
|  | S1 | 277 | 19 | 6.9 |
|  | S2 | 252 | 3 | 1.2 |

Prevalence by sex and child contact (small numbers)

| child\_contact\_u10 | Gender | N | N\_pos | Pct\_Pos |
| --- | --- | --- | --- | --- |
| 0 |  |  |  |  |
|  | F |  |  |  |
|  |  | 345 | 12 | 3.5 |
|  | M |  |  |  |
|  |  | 330 | 9 | 2.7 |
| 1 |  |  |  |  |
|  | F |  |  |  |
|  |  | 140 | 15 | 10.7 |
|  | M |  |  |  |
|  |  | 112 | 11 | 9.8 |
| 9,999 |  |  |  |  |
|  | F |  |  |  |
|  |  | 77 | 5 | 6.5 |
|  | M |  |  |  |
|  |  | 87 | 2 | 2.3 |

- Test with Fisher's Exact Test (p\<0.05)

**To be done, as requested by LG** - Descriptive analysis of NP serotype distribution PCV13 serotypes PPV23 serotypes NVT seroytpes - Correlations in serotype positivity among pairs - Correlations in serotype positivity among pairs over time (i.e., was there a likely transmission event where one adult in the pair was serotype positive and then the second adult in the pair became positive for the same serotype). - Comparison of genetic lineages for serotypes identified among pairs - UAD results – how will these be incorporated into the analysis?

### Regression models

Evaluate correlates of pneumococcal positivity.

* Multivariate logistic regression to identify relevant age groups

Assumes missing survey is no contact with kids…need to clean this up a bit (e.g., if survey totally missing exclude the observation)

mod1 <- glm(piab\_pos ~ child\_contact\_u12m + child\_contact\_13\_23m + child\_contact\_24\_59m + child\_contact\_5\_10y + child\_contact\_over10y , family='binomial', data=pcr\_survey\_mod)  
  
summary(mod1)

Test frequency of contact and N contacts age 1-10; each of which is associated in univariate analysis.

##   
## Call:  
## glm(formula = piab\_pos ~ child\_contact\_often + child\_contact\_24\_59m +   
## child\_contact\_5\_10y, family = "binomial", data = pcr\_survey\_mod)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.8180 -0.2563 -0.2415 -0.2415 2.7052   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.9546 0.6437 -4.590 4.44e-06 \*\*\*  
## child\_contact\_often2 0.6416 0.5912 1.085 0.2778   
## child\_contact\_often3 -0.6783 0.6587 -1.030 0.3031   
## child\_contact\_often999 -0.5658 0.6810 -0.831 0.4060   
## child\_contact\_24\_59m 1.1568 0.4615 2.507 0.0122 \*   
## child\_contact\_5\_10y 0.2332 0.4085 0.571 0.5680   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 371.87 on 926 degrees of freedom  
## Residual deviance: 342.47 on 921 degrees of freedom  
## AIC: 354.47  
##   
## Number of Fisher Scoring iterations: 6

Generalized estimating equation to account for correlation within individuals

mod1.gee <- gee(piab\_pos ~ child\_contact\_u12m+ child\_contact\_13\_23m + child\_contact\_24\_59m + child\_contact\_5\_10y + child\_contact\_over10y,id=pcr\_survey$ID, data=pcr\_survey\_mod, family=binomial(link='log'), corstr="exchangeable")

### Longitudinal data analysis

Use a Markov transition model with 2 states (colonized, uncolonized) to quantify acquisition and clearance rates and correlates of both. The model below estimates acquisition and clearance rates, and teh effect of recent contact with a child <12m on acquisition rates

mod.df <- pcr\_survey\_mod %>%  
 filter(child\_contact\_13\_23m!=9999 & !is.na(piab\_pos)) %>%  
 mutate(ID=as.factor(ID))  
  
mod.df$child\_contact\_u5 <-   
 mod.df$child\_contact\_u12m + mod.df$child\_contact\_13\_23m + mod.df$child\_contact\_24\_59m  
  
mod.df$child\_contact\_1\_10 <- mod.df$child\_contact\_13\_23m + mod.df$child\_contact\_24\_59m +  
mod.df$child\_contact\_5\_10y  
  
  
#N contacts  
mod.df$child\_contact\_1\_5 <- mod.df$child\_contact\_13\_23m + mod.df$child\_contact\_24\_59m  
  
  
l1 <- mod.df  
l1$day <- 0  
l1$day[l1$time==2] <- 14  
l1$day[l1$time==3] <- 14\*2  
l1$day[l1$time==4] <- 14\*3  
l1$day[l1$time==5] <- 14\*4  
l1$day[l1$time==6] <- 14\*5  
  
l1$state <- NA  
l1$state[l1$piab\_pos==0 & !is.na(l1$piab\_pos)] <- 1  
l1$state[l1$piab\_pos==1 & !is.na(l1$piab\_pos)] <- 2  
#l1$state <- as.factor(l1$state)  
  
  
q1 <- rbind(c( 0,0.01),  
 c(0.01,0)  
)  
  
l1 <- l1[order(l1$ID, l1$day),]  
  
l2 <- l1[, c('state','child\_contact\_u5','child\_contact\_1\_5','child\_contact','Gender','child\_contact\_u12m','child\_contact\_13\_23m',"child\_contact\_24\_59m","child\_contact\_5\_10y",'child\_contact\_over10y','child\_contact\_1\_10', 'ID','day', 'season','child\_contact\_often','child\_contact\_hours')] %>%  
 mutate(child\_contact\_2\_10y = if\_else( (child\_contact\_24\_59m==1 |child\_contact\_5\_10y==1) & child\_contact !=9999,1,  
 if\_else(child\_contact\_24\_59m==0 & child\_contact\_5\_10y==0 & child\_contact !=9999,0,   
 if\_else(child\_contact==9999 ,9999, 9999) ) ))  
  
msm.mod0 <- msm( state~ day , subject=ID , qmatrix=q1 , data=l2)  
msm.mod0

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1)  
##   
## Maximum likelihood estimates  
##   
## Transition intensities  
## Baseline   
## State 1 - State 1 -0.003369 (-0.005289,-0.002146)  
## State 1 - State 2 0.003369 ( 0.002146, 0.005289)  
## State 2 - State 1 0.056386 ( 0.035221, 0.090270)  
## State 2 - State 2 -0.056386 (-0.090270,-0.035221)  
##   
## -2 \* log-likelihood: 252.736

msm.mod1 <- msm( state~ day , subject=ID , qmatrix=q1 , data=l2,covariates = list("1-2" = ~ child\_contact\_13\_23m ))  
msm.mod1

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1, covariates = list(`1-2` = ~child\_contact\_13\_23m))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_13\_23m  
## State 1 - State 1 -0.003305 (-0.005221,-0.002092)   
## State 1 - State 2 0.003305 ( 0.002092, 0.005221) 2.139 (0.6294,7.273)  
## State 2 - State 1 0.056472 ( 0.035264, 0.090437) 1.000   
## State 2 - State 2 -0.056472 (-0.090437,-0.035264)   
##   
## -2 \* log-likelihood: 251.4985

msm.mod1 <- msm( state~ day , subject=ID , qmatrix=q1 , data=l2,covariates = list("1-2" = ~ child\_contact\_24\_59m ))  
msm.mod1

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1, covariates = list(`1-2` = ~child\_contact\_24\_59m))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_24\_59m  
## State 1 - State 1 -0.003375 (-0.005300,-0.002149)   
## State 1 - State 2 0.003375 ( 0.002149, 0.005300) 1.123 (0.2618,4.815)  
## State 2 - State 1 0.056457 ( 0.035240, 0.090449) 1.000   
## State 2 - State 2 -0.056457 (-0.090449,-0.035240)   
##   
## -2 \* log-likelihood: 252.7124

msm.mod1 <- msm( state~ day , subject=ID , qmatrix=q1 , data=l2,covariates = list("1-2" = ~ child\_contact\_5\_10y ))  
msm.mod1

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1, covariates = list(`1-2` = ~child\_contact\_5\_10y))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_5\_10y  
## State 1 - State 1 -0.003372 (-0.005311,-0.002141)   
## State 1 - State 2 0.003372 ( 0.002141, 0.005311) 1.505 (0.5565,4.07)  
## State 2 - State 1 0.056970 ( 0.035428, 0.091610) 1.000   
## State 2 - State 2 -0.056970 (-0.091610,-0.035428)   
##   
## -2 \* log-likelihood: 252.1362

msm.mod1 <- msm( state~ day , subject=ID , qmatrix=q1 , data=l2,covariates = list("1-2" = ~ child\_contact\_2\_10y ))  
msm.mod1

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1, covariates = list(`1-2` = ~child\_contact\_2\_10y))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_2\_10y  
## State 1 - State 1 -0.003364 (-0.005304,-0.002134)   
## State 1 - State 2 0.003364 ( 0.002134, 0.005304) 1 (0.9999,1)   
## State 2 - State 1 0.057433 ( 0.035475, 0.092982) 1   
## State 2 - State 2 -0.057433 (-0.092982,-0.035475)   
##   
## -2 \* log-likelihood: 251.6802

msm.mod1 <- msm( state~ day , subject=ID , qmatrix=q1 , data=l2,covariates = list("1-2" = ~ child\_contact\_13\_23m + child\_contact\_24\_59m + child\_contact\_5\_10y ))  
msm.mod1

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1, covariates = list(`1-2` = ~child\_contact\_13\_23m + child\_contact\_24\_59m + child\_contact\_5\_10y))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_13\_23m  
## State 1 - State 1 -0.003296 (-0.005227,-0.002078)   
## State 1 - State 2 0.003296 ( 0.002078, 0.005227) 2.147 (0.6056,7.609)  
## State 2 - State 1 0.056876 ( 0.035403, 0.091372) 1.000   
## State 2 - State 2 -0.056876 (-0.091372,-0.035403)   
## child\_contact\_24\_59m child\_contact\_5\_10y   
## State 1 - State 1   
## State 1 - State 2 0.7742 (0.1607,3.73) 1.507 (0.5283,4.296)  
## State 2 - State 1 1.0000 1.000   
## State 2 - State 2   
##   
## -2 \* log-likelihood: 250.9393

msm.mod1b <- msm( state~ day , subject=ID , qmatrix=q1 , data=l2,covariates = list("1-2" = ~ child\_contact\_1\_5))  
msm.mod1b

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1, covariates = list(`1-2` = ~child\_contact\_1\_5))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_1\_5   
## State 1 - State 1 -0.003354 (-0.005282,-0.002130)   
## State 1 - State 2 0.003354 ( 0.002130, 0.005282) 1.442 (0.6293,3.304)  
## State 2 - State 1 0.056675 ( 0.035336, 0.090899) 1.000   
## State 2 - State 2 -0.056675 (-0.090899,-0.035336)   
##   
## -2 \* log-likelihood: 252.0795

msm.mod1b <- msm( state~ day , subject=ID , qmatrix=q1 , data=l2,covariates = list("1-2" = ~ child\_contact\_u5))  
msm.mod1b

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1, covariates = list(`1-2` = ~child\_contact\_u5))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_u5   
## State 1 - State 1 -0.003359 (-0.005287,-0.002135)   
## State 1 - State 2 0.003359 ( 0.002135, 0.005287) 1.279 (0.6644,2.461)  
## State 2 - State 1 0.056647 ( 0.035326, 0.090836) 1.000   
## State 2 - State 2 -0.056647 (-0.090836,-0.035326)   
##   
## -2 \* log-likelihood: 252.245

msm.mod1c <- msm( state~ day , subject=ID , qmatrix=q1 , data=l2,covariates = list("1-2" = ~ child\_contact\_1\_10))  
msm.mod1c

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1, covariates = list(`1-2` = ~child\_contact\_1\_10))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_1\_10   
## State 1 - State 1 -0.003361 (-0.005303,-0.002130)   
## State 1 - State 2 0.003361 ( 0.002130, 0.005303) 1.35 (0.7713,2.364)  
## State 2 - State 1 0.057089 ( 0.035479, 0.091860) 1.00   
## State 2 - State 2 -0.057089 (-0.091860,-0.035479)   
##   
## -2 \* log-likelihood: 251.7521

#clearance rate as a function of child contacts?  
msm.mod1e <- msm( state~ day , subject=ID , qmatrix=q1 , data=l2,covariates = list( "2-1" = ~ child\_contact\_1\_10))  
msm.mod1e

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1, covariates = list(`2-1` = ~child\_contact\_1\_10))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_1\_10   
## State 1 - State 1 -0.003868 (-0.006301,-0.002375)   
## State 1 - State 2 0.003868 ( 0.002375, 0.006301) 1.0000   
## State 2 - State 1 0.075442 ( 0.045267, 0.125733) 0.5836 (0.3382,1.007)  
## State 2 - State 2 -0.075442 (-0.125733,-0.045267)   
##   
## -2 \* log-likelihood: 248.6498

Frequency of child contact and acquisition and clearance rates. Those with contact with children have more frequent acquisition. Not clear if can identify acquistiion and clearance rates from these data, especially without serotype info.

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l3, qmatrix = q1)  
##   
## Maximum likelihood estimates  
##   
## Transition intensities  
## Baseline   
## State 1 - State 1 -0.003465 (-0.005623,-0.002135)  
## State 1 - State 2 0.003465 ( 0.002135, 0.005623)  
## State 2 - State 1 0.068505 ( 0.042067, 0.111560)  
## State 2 - State 2 -0.068505 (-0.111560,-0.042067)  
##   
## -2 \* log-likelihood: 227.6195

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l3, qmatrix = q1, covariates = list(`1-2` = ~child\_contact\_often1 + child\_contact\_often2))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_often1  
## State 1 - State 1 -0.00305 (-0.005237,-0.001776)   
## State 1 - State 2 0.00305 ( 0.001776, 0.005237) 3.718 (1.499,9.223)   
## State 2 - State 1 0.07115 ( 0.043016, 0.117699) 1.000   
## State 2 - State 2 -0.07115 (-0.117699,-0.043016)   
## child\_contact\_often2  
## State 1 - State 1   
## State 1 - State 2 0.8304 (0.1786,3.86)  
## State 2 - State 1 1.0000   
## State 2 - State 2   
##   
## -2 \* log-likelihood: 218.9552

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l3, qmatrix = q1, covariates = list(`1-2` = ~child\_contact\_often1, `2-1` = ~child\_contact\_often1))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_often1   
## State 1 - State 1 -0.003485 (-0.006334,-0.001918)   
## State 1 - State 2 0.003485 ( 0.001918, 0.006334) 2.7205 (0.9910,7.469)  
## State 2 - State 1 0.090627 ( 0.050433, 0.162854) 0.5019 (0.1841,1.369)  
## State 2 - State 2 -0.090627 (-0.162854,-0.050433)   
##   
## -2 \* log-likelihood: 217.2109

or hours of contact. acquisition rate is lower for those with 4-8 and 8+ hours of reported contact–these people tend to have less frequent contacts

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1, covariates = list(`1-2` = ~child\_contact\_hours1 + child\_contact\_hours2))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_hours1  
## State 1 - State 1 -0.00323 (-0.005184,-0.002013)   
## State 1 - State 2 0.00323 ( 0.002013, 0.005184) 2.426 (1.005,5.858)   
## State 2 - State 1 0.05779 ( 0.035767, 0.093390) 1.000   
## State 2 - State 2 -0.05779 (-0.093390,-0.035767)   
## child\_contact\_hours2  
## State 1 - State 1   
## State 1 - State 2 1.243 (0.3404,4.54)   
## State 2 - State 1 1.000   
## State 2 - State 2   
##   
## -2 \* log-likelihood: 248.9193

It appears those with shorter contacts have higher carriage than those with longer contacts. This could be a) confounding by age of contacts b) confoduning by frequency of contacts c) some boosting effect from frequent/intense contact. Descriptive analyses suggest it is (b)

Sems that those with contact with 1-10 year olds with lower duration contact have highest estimate for acquisition rate

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2.cat, qmatrix = q1, covariates = list(`1-2` = ~child\_contact\_cat))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact\_cat1   
## State 1 - State 1 -0.003124 (-0.005039,-0.001937)   
## State 1 - State 2 0.003124 ( 0.001937, 0.005039) 2.635 (0.9757,7.115)  
## State 2 - State 1 0.054488 ( 0.033412, 0.088860) 1.000   
## State 2 - State 2 -0.054488 (-0.088860,-0.033412)   
## child\_contact\_cat2 child\_contact\_cat3   
## State 1 - State 1   
## State 1 - State 2 2.195 (0.7326,6.579) 1.988 (0.4262,9.27)  
## State 2 - State 1 1.000 1.000   
## State 2 - State 2   
##   
## -2 \* log-likelihood: 246.1236

##   
## 1 2  
## 0 0.97765363 0.02234637  
## 1 0.93167702 0.06832298  
## 2 0.88157895 0.11842105  
## 3 0.92753623 0.07246377

##   
## S1 S2  
## 0 0.58620690 0.58264463  
## 1 0.23448276 0.12190083  
## 2 0.13333333 0.19421488  
## 3 0.04597701 0.10123967

dichotomized contacts age 1-10y

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2.cat, qmatrix = q1, covariates = list(`1-2` = ~child\_contact1\_10\_dic))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline child\_contact1\_10\_dic  
## State 1 - State 1 -0.003313 (-0.005221,-0.002103)   
## State 1 - State 2 0.003313 ( 0.002103, 0.005221) 1.552 (0.6361,3.789)   
## State 2 - State 1 0.054195 ( 0.033321, 0.088144) 1.000   
## State 2 - State 2 -0.054195 (-0.088144,-0.033321)   
##   
## -2 \* log-likelihood: 249.5345

Does duration or acquisition rate vary by sex? No.

Effect of positive household contact. This quantifies transmission within households.

mod.df <- pcr\_survey %>%  
 filter( !is.na(piab\_pos)) %>%  
 mutate(ID=as.factor(ID))  
  
recent\_hh\_col <- mod.df %>%  
 group\_by(Household,time ) %>%  
 summarize(piab\_pos=max(piab\_pos)) %>%  
 mutate(time=time+1) %>%  
 rename(piab\_pos\_HH\_prev = piab\_pos)  
  
recent\_ID\_col <- mod.df %>%  
 group\_by(Household,ID,time ) %>%  
 summarize(piab\_pos=sum(piab\_pos)) %>%  
 mutate(time=time+1) %>%  
 rename(piab\_pos\_ID\_prev = piab\_pos)  
  
#colonized\_partner\_prev is defined based on a positive household without positive individual on previous time  
recent\_col <- merge(recent\_hh\_col,recent\_ID\_col, by=c('Household', 'time') ) %>%  
 mutate(colonized\_partner\_prev = piab\_pos\_HH\_prev - piab\_pos\_ID\_prev)  
  
mod.df <- merge(mod.df,recent\_col, by=c('ID','Household', 'time'))   
  
#N contacts  
mod.df$child\_contact\_1\_10 <- mod.df$child\_contact\_13\_23m + mod.df$child\_contact\_24\_59m +mod.df$child\_contact\_5\_10y  
  
  
l1 <- mod.df  
l1$day <- 0  
l1$day[l1$time==2] <- 14  
l1$day[l1$time==3] <- 14\*2  
l1$day[l1$time==4] <- 14\*3  
l1$day[l1$time==5] <- 14\*4  
l1$day[l1$time==6] <- 14\*5  
  
l1$state <- NA  
l1$state[l1$piab\_pos==0 & !is.na(l1$piab\_pos)] <- 1  
l1$state[l1$piab\_pos==1 & !is.na(l1$piab\_pos)] <- 2  
#l1$state <- as.factor(l1$state)  
  
  
q1 <- rbind(c( 0,0.01),  
 c(0.01,0)  
)  
  
l1 <- l1[order(l1$ID, l1$day),]  
  
l2 <- l1[, c('state','Gender','child\_contact\_1\_10', 'ID','day','colonized\_partner\_prev')]   
  
table(l2$state, l2$colonized\_partner\_prev)  
  
msm.mod1 <- msm( state~ day , subject=ID , qmatrix=q1 , data=l2,covariates = list("1-2" = ~ colonized\_partner\_prev ))  
  
msm.mod1

as a test, try same analysis using lyta as the outcome. This is problematic because we cannot distrinuish lyta strains, but might give some hint.

##   
## 0 1  
## 1 562 101  
## 2 221 18

## Warning in msm.check.times(mf$"(time)", mf$"(subject)", mf$"(state)"): Subject  
## S2\_84 only has one complete observation, which doesn't give any information

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1)  
##   
## Maximum likelihood estimates  
##   
## Transition intensities  
## Baseline   
## State 1 - State 1 -0.009295 (-0.01222,-0.00707)  
## State 1 - State 2 0.009295 ( 0.00707, 0.01222)  
## State 2 - State 1 0.024349 ( 0.01837, 0.03228)  
## State 2 - State 2 -0.024349 (-0.03228,-0.01837)  
##   
## -2 \* log-likelihood: 570.8521

## Warning in msm.check.times(mf$"(time)", mf$"(subject)", mf$"(state)"): Subject  
## S2\_84 only has one complete observation, which doesn't give any information

##   
## Call:  
## msm(formula = state ~ day, subject = ID, data = l2, qmatrix = q1, covariates = list(`1-2` = ~colonized\_partner\_prev))  
##   
## Maximum likelihood estimates  
## Baselines are with covariates set to their means  
##   
## Transition intensities with hazard ratios for each covariate  
## Baseline colonized\_partner\_prev  
## State 1 - State 1 -0.009278 (-0.012208,-0.007051)   
## State 1 - State 2 0.009278 ( 0.007051, 0.012208) 1.071 (0.5219,2.198)   
## State 2 - State 1 0.024338 ( 0.018359, 0.032265) 1.000   
## State 2 - State 2 -0.024338 (-0.032265,-0.018359)   
##   
## -2 \* log-likelihood: 570.8185

## Notes and challenges

* The analyses will all be done initially based on piaB positivity. Given the low rate of carriage in adults, this is probably fine to assume that if a person is positive at multiple time points, or multiple people in the household are positive, it is the same serotype. (the former has been confirmed–we have 1 person positive for 15B/C at all time points). Once serotype data are available, we can more precisely define duration and transmission, if there is discordance within individuals or households
* The evaluations of household transmission are likely underpowered due to low levels of carriage and the fact that when one household member is positive, the other individual is often also positive at the same time point.