class15

Rodeon Malinovski

Background

Pertussis (more commonly known as *whooping cough*) is a highly contagious respiratory disease caused by the bacterium **Bordetella pertussis**. B. **pertussis** attacks cells lining the airways. In a pertussis infection, the bacteria use adhesive proteins to stick to lining cells whilst releasing toxins that damage the cells, trigger inflammation and increase mucus production leading to uncontrollable violent coughing.

1. Investigating Pertussis Cases by Year

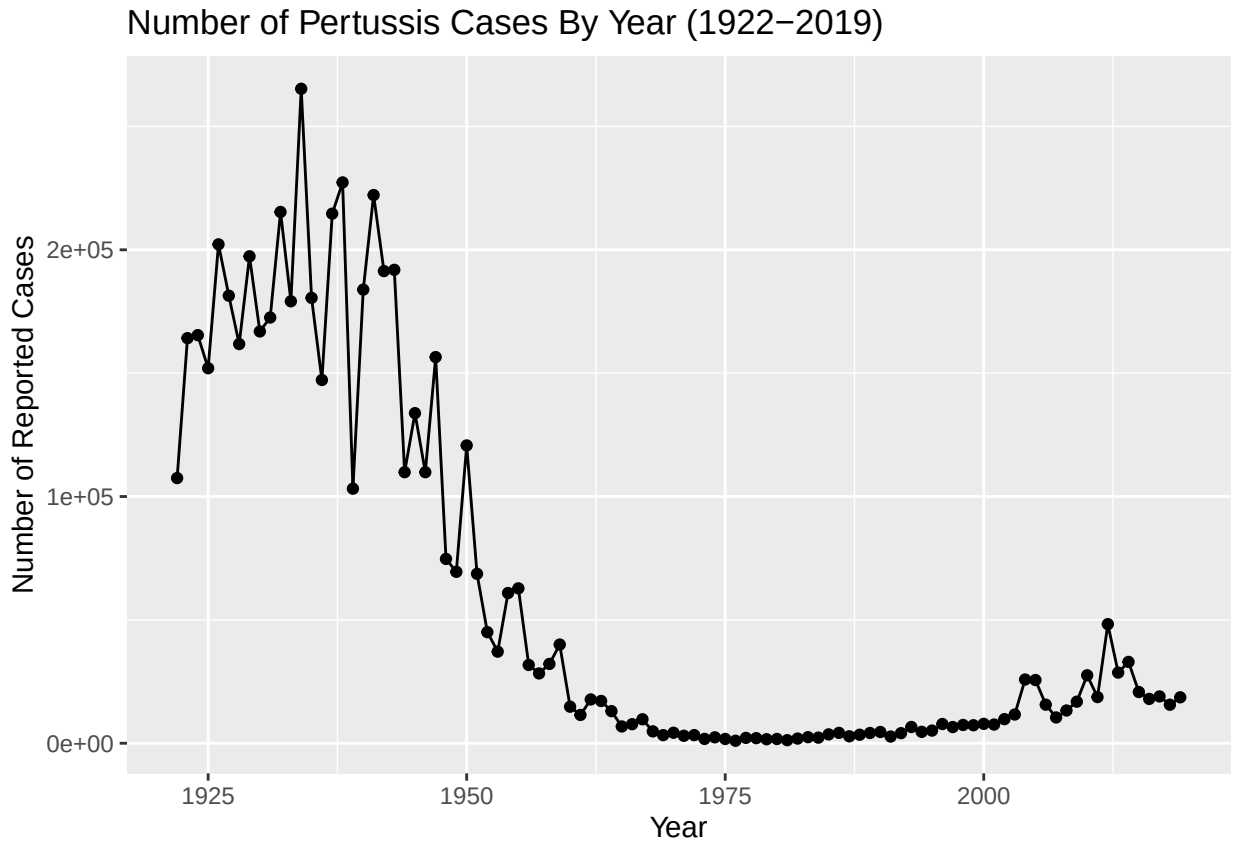
```
library(datapasta)
```

```
#For format, copy data from website, then click 'Addins > Paste as data.frame' in RStudio
cdc <- data.frame(
                                  Year = c(1922L,1923L,1924L,1925L,
                                           1926L,1927L,1928L,1929L,1930L,1931L,
                                           1932L,1933L,1934L,1935L,1936L,
                                           1937L,1938L,1939L,1940L,1941L,1942L,
                                           1943L,1944L,1945L,1946L,1947L,
                                           1948L,1949L,1950L,1951L,1952L,
                                           1953L,1954L,1955L,1956L,1957L,1958L,
                                           1959L,1960L,1961L,1962L,1963L,
                                           1964L,1965L,1966L,1967L,1968L,1969L,
                                           1970L,1971L,1972L,1973L,1974L,
                                           1975L,1976L,1977L,1978L,1979L,1980L,
                                           1981L,1982L,1983L,1984L,1985L,
                                           1986L,1987L,1988L,1989L,1990L,
                                           1991L,1992L,1993L,1994L,1995L,1996L,
                                           1997L,1998L,1999L,2000L,2001L,
                                           2002L,2003L,2004L,2005L,2006L,2007L,
                                           2008L,2009L,2010L,2011L,2012L,
                                           2013L,2014L,2015L,2016L,2017L,2018L,
                                           2019L),
        No..Reported.Pertussis.Cases = c(107473,164191,165418,152003,
                                           202210,181411,161799,197371,
                                           166914,172559,215343,179135,265269,
                                           180518,147237,214652,227319,103188,
                                           183866,222202,191383,191890,109873,
                                           133792,109860,156517,74715,69479,
                                           120718,68687,45030,37129,60886,
                                           62786,31732,28295,32148,40005,
```

```
                                    14809,11468,17749,17135,13005,6799,
                                    7717,9718,4810,3285,4249,3036,
                                    3287,1759,2402,1738,1010,2177,2063,
                                    1623,1730,1248,1895,2463,2276,
                                    3589,4195,2823,3450,4157,4570,
                                    2719,4083,6586,4617,5137,7796,6564,
                                    7405,7298,7867,7580,9771,11647,
                                    25827,25616,15632,10454,13278,
                                    16858,27550,18719,48277,28639,32971,
                                    20762,17972,18975,15609,18617)
        )
```

Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```
library(ggplot2)
ggplot(cdc) + aes(x = Year, y = No..Reported.Pertussis.Cases) + geom_point() + geom_line() +
  labs(title = "Number of Pertussis Cases By Year (1922-2019)", x = "Year", y = "Number of Reported Cas
```



2. Comparing Two Vaccines (wP & aP)

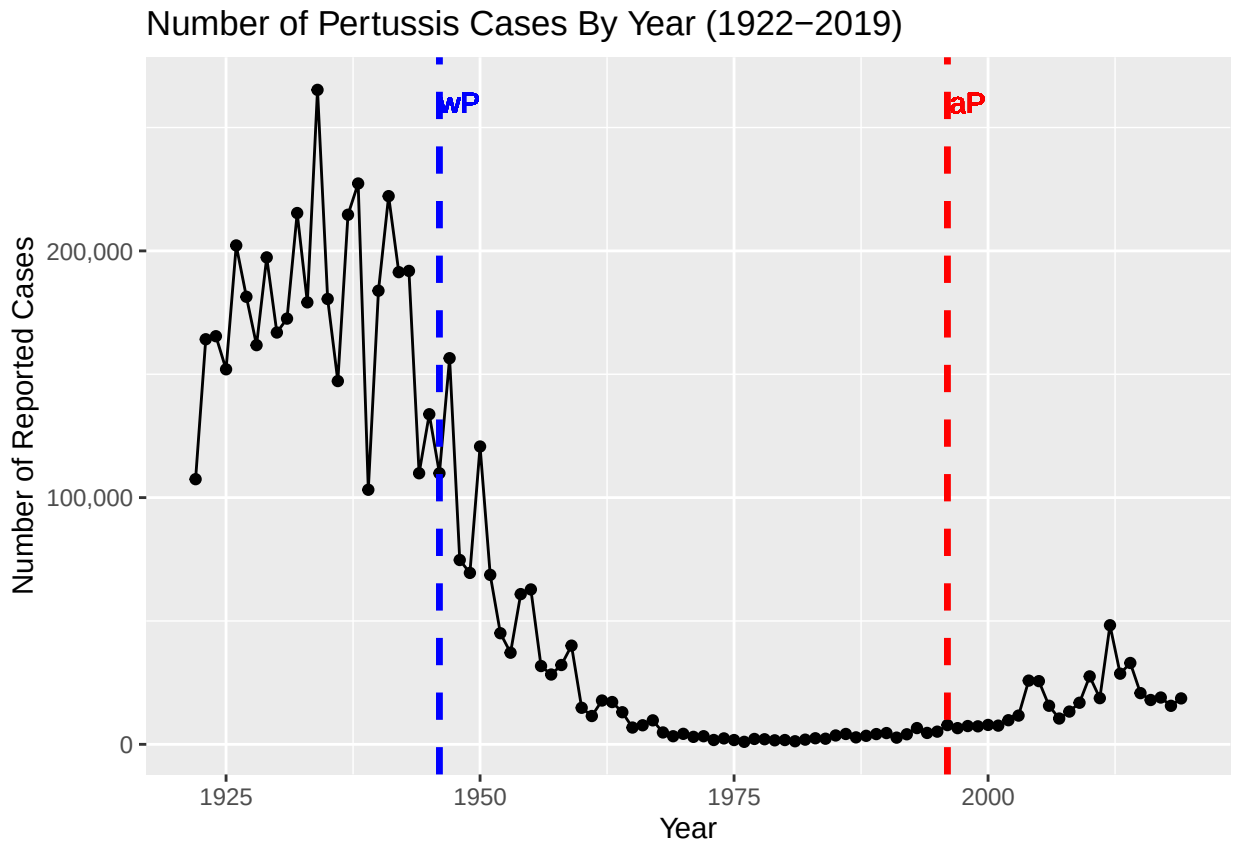
wP: Traditional whole-cell pertussis vaccine (killed bacteria cells presented to immune system) - introduced 1946 ap: Acellular pertussis vaccine (only parts of the cell deemed most important for identification presented to the immune system) helps to mediate adverse reactions to injecting whole bacterial cells - introduced 1996

Q2. Using the ggplot geom_vline() function add lines to your previous plot for the 1946 introduction of the wP vaccine and the 1996 switch to aP vaccine (see example in the hint below). What do you notice? *The cases go up shortly after the introduction of the new vaccine.*

```
require(scales)
```

```
## Loading required package: scales
```

```
ggplot(cdc) + aes(x = Year, y = No..Reported.Pertussis.Cases) + geom_point() + geom_line() +
  labs(title = "Number of Pertussis Cases By Year (1922-2019)", x = "Year", y = "Number of Reported Cas
```



Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend? *I thought that the reason for the case rates rising would be because the new vaccine is less effective, and the delay is because of heard immunity due to the old vaccine, but the pertussis field has several different hypotheses for the resurgence of pertussis including (in no particular order): 1) more sensitive PCR-based testing, 2) vaccination hesitancy 3) bacterial evolution (escape from vaccine immunity), 4) waning of immunity in adolescents originally primed as infants with the newer aP vaccine as compared to the older wP vaccine.*

3. Exploring CMI-PB Data

Why is this vaccine-preventable disease on the upswing? The new and ongoing CMI-PB project aims to provide the scientific community with this very information. In particular, CMI-PB tracks and makes freely

available long-term humoral and cellular immune response data for a large number of individuals who received either DTwP or DTaP combination vaccines in infancy followed by Tdap booster vaccinations. This includes complete API access to longitudinal RNA-Seq, AB Titer, Olink, and live cell assay results directly from their website: https://www.cmi-pb.org/

```
#The CMI-PB API (like most APIs) sends responses in JSON format
library(jsonlite)
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)
head(subject, 3)
```

```
##   subject_id infancy_vac biological_sex              ethnicity  race
## 1          1          wP         Female Not Hispanic or Latino White
## 2          2          wP         Female Not Hispanic or Latino White
## 3          3          wP         Female                 Unknown White
##   year_of_birth date_of_boost   study_name
## 1    1986-01-01    2016-09-12 2020_dataset
## 2    1968-01-01    2019-01-28 2020_dataset
## 3    1983-01-01    2016-10-10 2020_dataset
```

```
table(subject$infancy_vac)
```

```
##
## aP wP
## 47 49
```

```
table(subject$biological_sex)
```

```
##
## Female   Male
##     66     30
```

```
table(subject$race)
```

```
##
##             American Indian/Alaska Native
##                                          1
##                                      Asian
##                                         27
##                 Black or African American
##                                          2
##                          More Than One Race
##                                         10
## Native Hawaiian or Other Pacific Islander
##                                          2
##                   Unknown or Not Reported
##                                         14
##                                      White
##                                         40
```

Q4. How may aP and wP infancy vaccinated subjects are in the dataset? *There are 47 people vaccinated with aP and 49 people vaccinated with wP.*

Q5. How many Male and Female subjects/patients are in the dataset? *66 females, 30 males*

Q6. What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc...)? *American Indian/Alaska Native: 1, Asian: 27, Black or African American: 2, More Than One Race: 10, Native Hawaiian or Other Pacific Islander: 2, Unknown or Not Reported: 14, White: 40*

Working with Dates

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
aP_dates <- subject %>% filter(infancy_vac == "aP")
wP_dates <- subject %>% filter(infancy_vac == "wP")
aP_ages <- time_length( today() - ymd(aP_dates$year_of_birth),  "years")
wP_ages <- time_length( today() - ymd(wP_dates$year_of_birth),  "years")

mean(wP_ages)
```

```
## [1] 35.34431
```

```
mean(aP_ages)
```

```
## [1] 24.49986
```

```
t.test(wP_ages, aP_ages)
```

```
##
##  Welch Two Sample t-test
##
## data:  wP_ages and aP_ages
## t = 12.092, df = 51.082, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.044045 12.644857
## sample estimates:
## mean of x mean of y
##  35.34431  24.49986
```

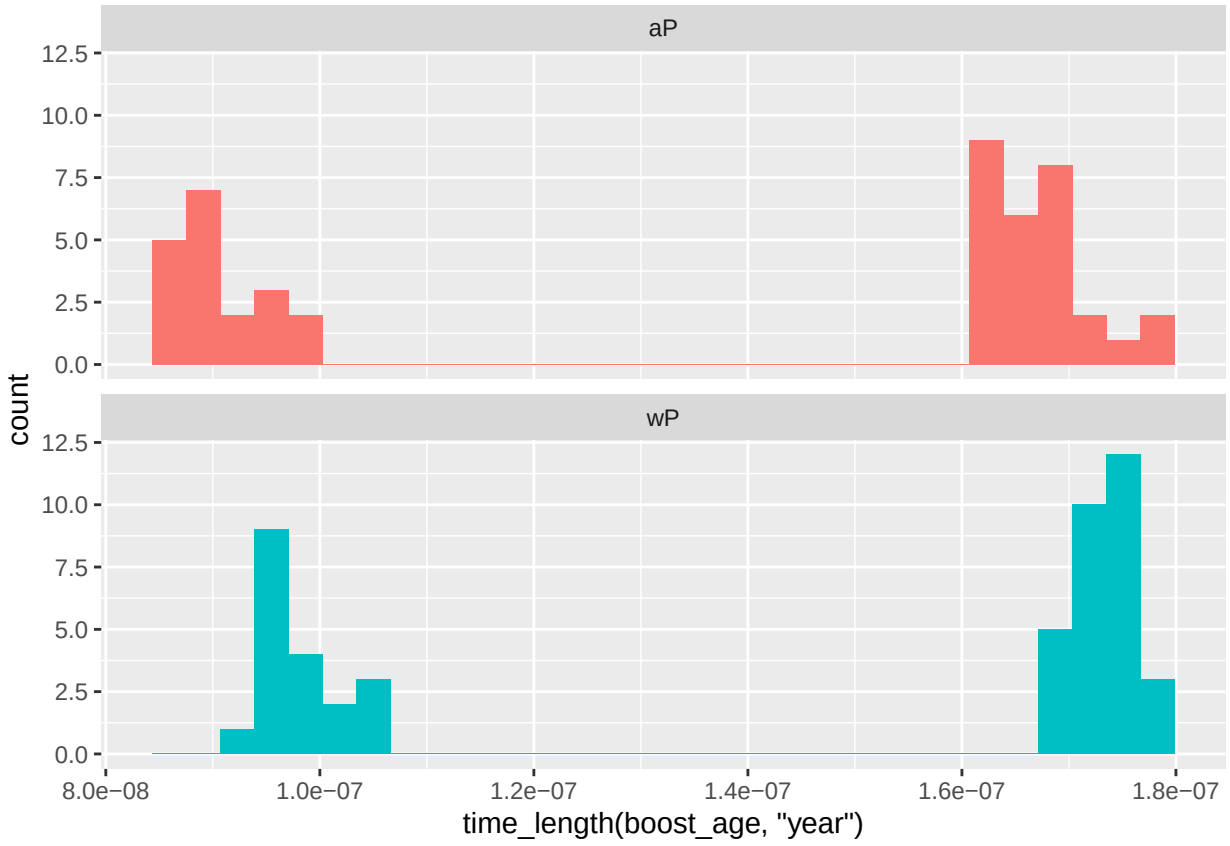
```
boost_age <- time_length( today() - ymd(subject$date_of_boost), "years")
```

Q7. Using this approach determine (i) the average age of wP individuals, (ii) the average age of aP individuals; and (iii) are they significantly different? *Average age of wP: 35 y.o.; Average age of aP: 24 y.o.; yes, they are significantly different (p-value < 2.2e-16).*

Q8. Determine the age of all individuals at time of boost? *Results stored in boost_age*

```
ggplot(subject) +
  aes(time_length(boost_age, "year"),
      fill=as.factor(infancy_vac)) +
  geom_histogram(show.legend=FALSE) +
  facet_wrap(vars(infancy_vac), nrow=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Q9. With the help of a faceted boxplot (see below), do you think these two groups are significantly different? *Yes, the two groups are different in their age. wP is pretty evenly distributed while aP is hevily skewed in favor of younger people.*

Joining Datasets

Q9. Complete the code to join specimen and subject tables to make a new merged data frame containing all specimen records along with their associated subject details: *v*

Q10. Now using the same procedure join meta with titer data so we can further analyze this data in terms of time of visit aP/wP, male/female etc. *v*

```
library(dplyr)
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector = TRUE)
titer <- read_json("https://www.cmi-pb.org/api/ab_titer", simplifyVector = TRUE)

meta <- inner_join(specimen, subject)
```

```
## Joining, by = "subject_id"
```

```
dim(meta)
```

```
## [1] 729  13
```

```
head(meta)
```

```
##   specimen_id subject_id actual_day_relative_to_boost
## 1           1          1                           -3
## 2           2          1                          736
## 3           3          1                            1
## 4           4          1                            3
## 5           5          1                            7
## 6           6          1                           11
##   planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
## 1                             0         Blood     1          wP         Female
## 2                           736         Blood    10          wP         Female
## 3                             1         Blood     2          wP         Female
## 4                             3         Blood     3          wP         Female
## 5                             7         Blood     4          wP         Female
## 6                            14         Blood     5          wP         Female
##                ethnicity  race year_of_birth date_of_boost   study_name
## 1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
## 2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
## 3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
## 4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
## 5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
## 6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
```

```
abdata <- inner_join(meta, titer)
```

```
## Joining, by = "specimen_id"
```

```
dim(abdata)
```

```
## [1] 32675    19
```

```
head(abdata)
```

```
##   specimen_id subject_id actual_day_relative_to_boost
## 1           1          1                           -3
## 2           1          1                           -3
## 3           1          1                           -3
## 4           1          1                           -3
## 5           1          1                           -3
## 6           1          1                           -3
##   planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
## 1                             0         Blood     1          wP         Female
## 2                             0         Blood     1          wP         Female
## 3                             0         Blood     1          wP         Female
## 4                             0         Blood     1          wP         Female
## 5                             0         Blood     1          wP         Female
## 6                             0         Blood     1          wP         Female
##                ethnicity  race year_of_birth date_of_boost   study_name isotype
## 1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset     IgE
## 2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset     IgE
```

```
## 3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset     IgG
## 4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset     IgG
## 5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset     IgG
## 6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset     IgE
##   is_antigen_specific antigen   ab_titer  unit lower_limit_of_detection
## 1               FALSE   Total 1110.21154 UG/ML                     NaN
## 2               FALSE   Total 2708.91616 IU/ML               29.170000
## 3                TRUE      PT   68.56614 IU/ML                0.530000
## 4                TRUE     PRN  332.12718 IU/ML                1.070000
## 5                TRUE     FHA 1887.12263 IU/ML                0.064000
## 6                TRUE     ACT    0.10000 IU/ML                2.816431
```

```
table(abdata$isotype)
```

```
##
##  IgE  IgG IgG1 IgG2 IgG3 IgG4
## 6698 1413 6141 6141 6141 6141
```

```
table(abdata$visit)
```

```
##
##    1    2    3    4    5    6    7    8
## 5795 4640 4640 4640 4640 4320 3920   80
```

Q11. How many specimens (i.e. entries in abdata) do we have for each isotype? *IgE: 6698, IgG: 1413, IgG1: 6141, IgG2: 6141, IgG3: 6141, IgG4: 6141*

Q12. What do you notice about the number of visit 8 specimens compared to other visits? *It's much lower than the numbers of other visits - the project is ongoing and the data is still being collected (visit 8 being the most recent of the visits).*

4. Examine IgG1 Ab Titer Levels

Here we exclude the incomplete visit 8 data...

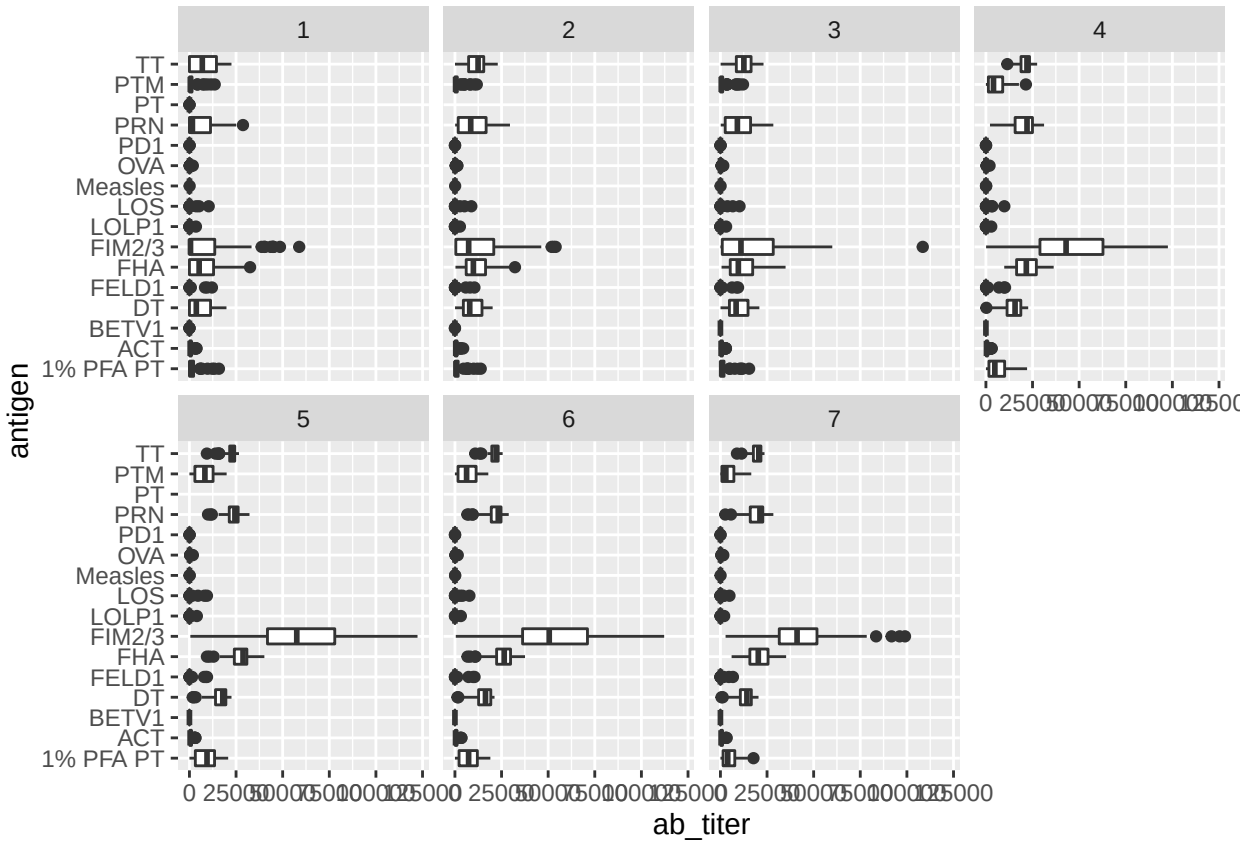
```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

```
##   specimen_id subject_id actual_day_relative_to_boost
## 1           1          1                           -3
## 2           1          1                           -3
## 3           1          1                           -3
## 4           1          1                           -3
## 5           1          1                           -3
## 6           1          1                           -3
##   planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
## 1                             0         Blood     1          wP         Female
## 2                             0         Blood     1          wP         Female
## 3                             0         Blood     1          wP         Female
## 4                             0         Blood     1          wP         Female
```

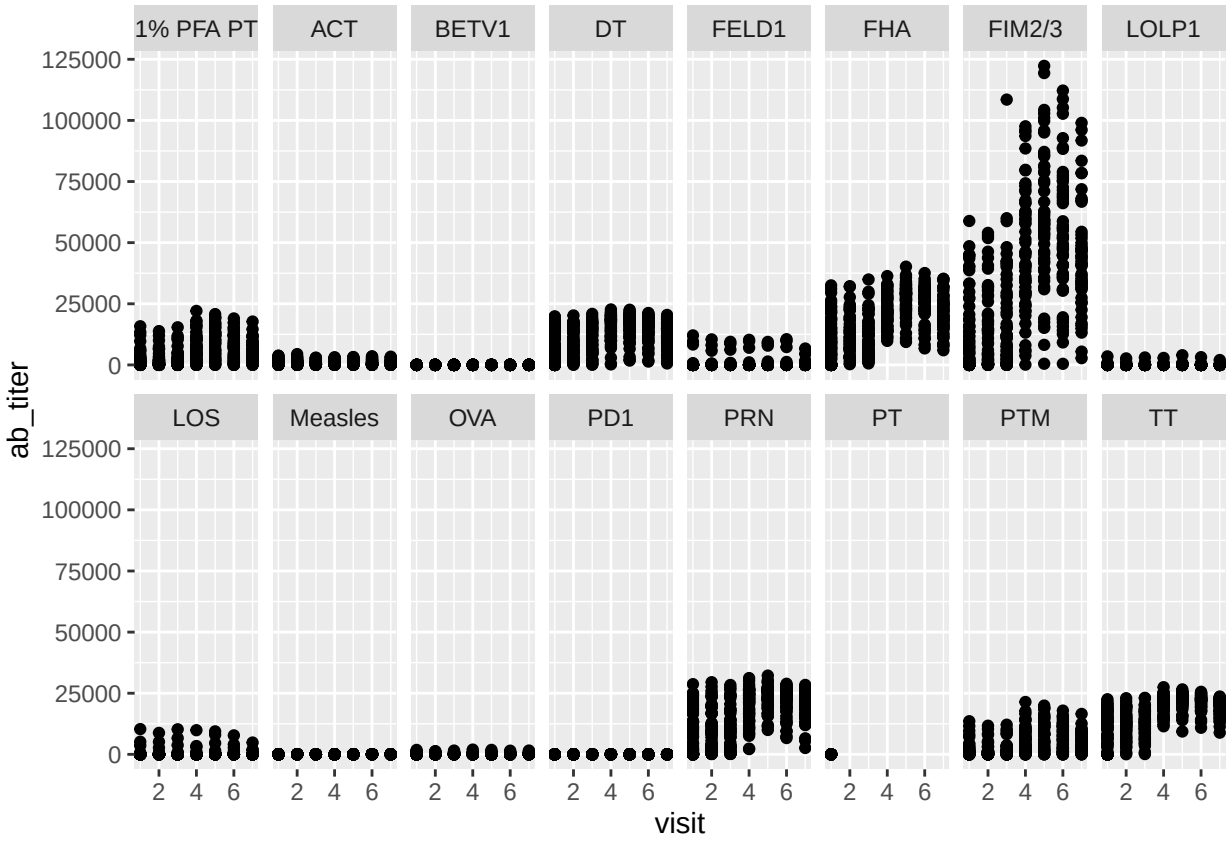
```
## 5                              0         Blood     1          wP         Female
## 6                              0         Blood     1          wP         Female
##                ethnicity  race year_of_birth date_of_boost   study_name isotype
## 1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset    IgG1
## 2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset    IgG1
## 3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset    IgG1
## 4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset    IgG1
## 5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset    IgG1
## 6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset    IgG1
##   is_antigen_specific antigen   ab_titer  unit lower_limit_of_detection
## 1                TRUE     ACT 274.355068 IU/ML                 3.848750
## 2                TRUE     LOS  10.974026 IU/ML                 4.357917
## 3                TRUE   FELD1   1.448796 IU/ML                 2.699944
## 4                TRUE   BETV1   0.100000 IU/ML                 1.734784
## 5                TRUE   LOLP1   0.100000 IU/ML                 2.550606
## 6                TRUE Measles  36.277417 IU/ML                 4.438966
```

Q13. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:

```
ggplot(ig1) +
  aes(ab_titer, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```

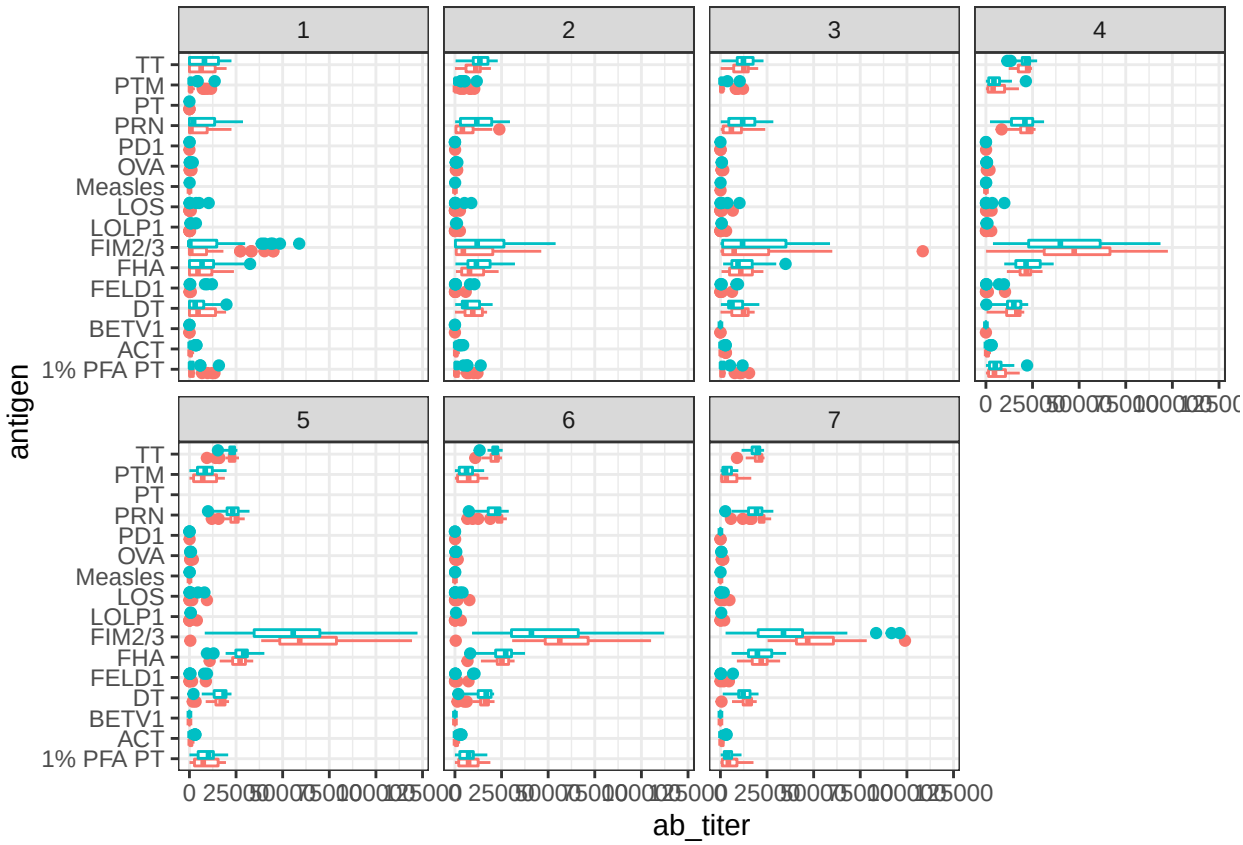


```
ggplot(ig1) +
  aes(visit, ab_titer) +
  geom_point() +
  facet_wrap(vars(antigen), nrow=2)
```



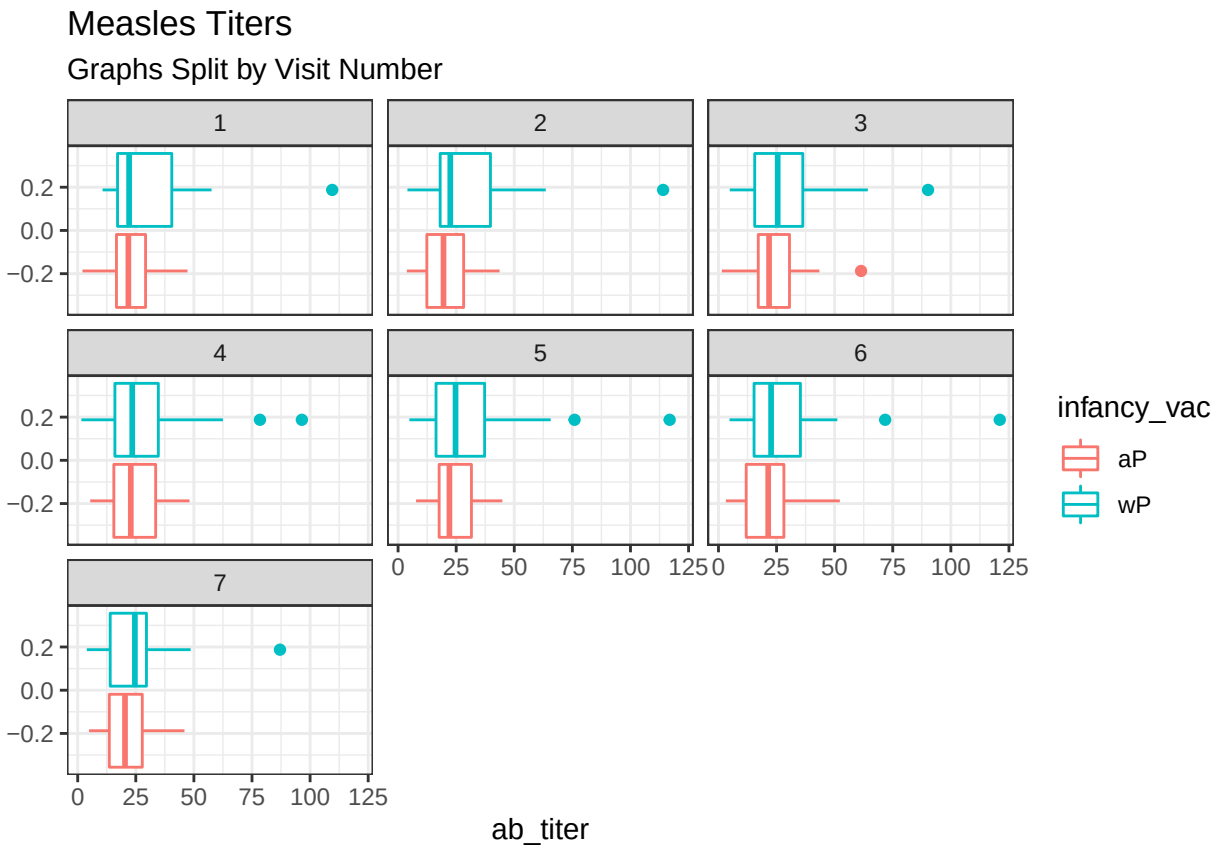
Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others? *The FIM2/3 antigen shows the greatest difference over time, probably because it is involved with pertussis*

```
ggplot(ig1) +
  aes(ab_titer, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  theme_bw()
```

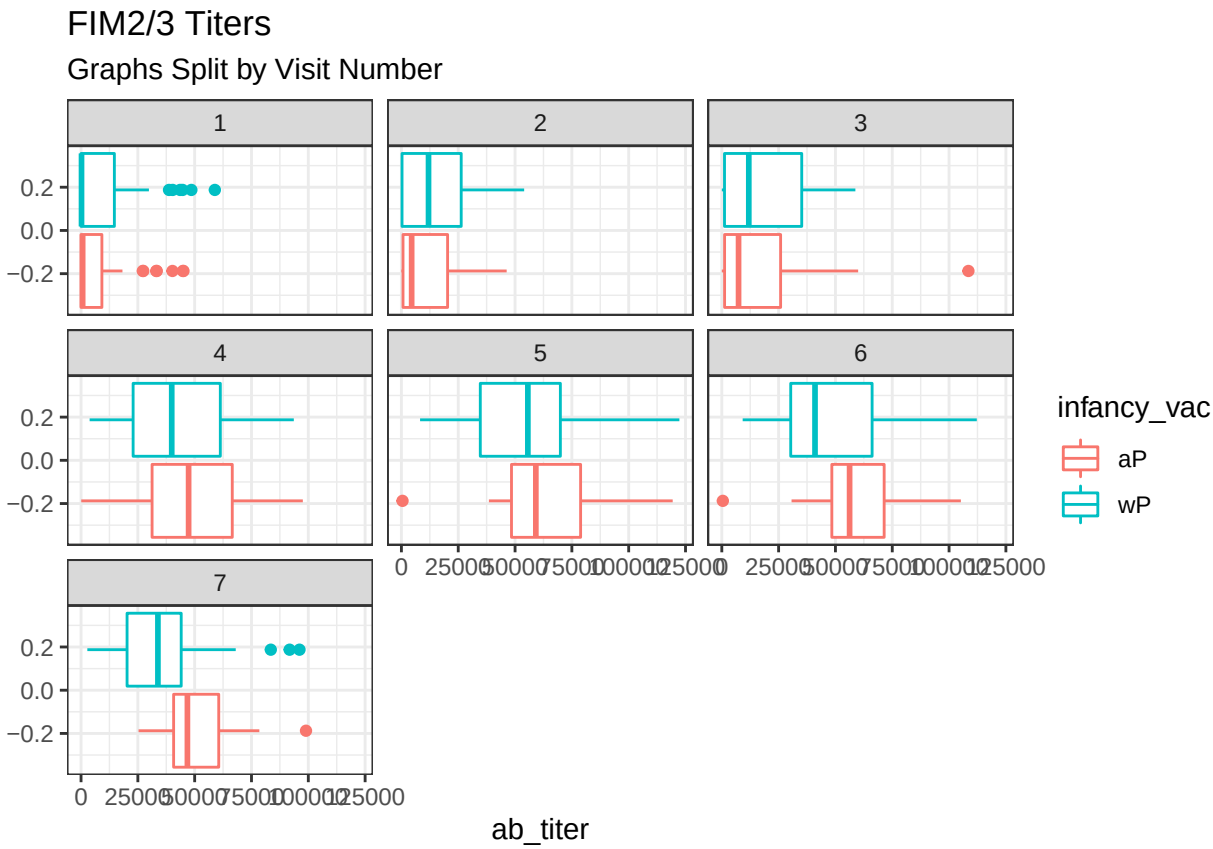


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. I will use FIM2/3 and control measles as well.

```
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(ab_titer, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw() + labs(title="Measles Titers", subtitle="Graphs Split by Visit Number")
```




```
filter(ig1, antigen=="FIM2/3") %>%
  ggplot() +
  aes(ab_titer, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw() + labs(title="FIM2/3 Titers", subtitle="Graphs Split by Visit Number")
```



Q16. What do you notice about these two antigens time course and the FIM2/3 data in particular? *FIM2/3 titers ride rapidly after receiving both vaccinations and then starts falling after visit 5, more rapidly in wP patients.*

Q17. Do you see any clear difference in aP vs. wP responses? *Yes, not necessarily in the rise, but the wP response diminishes faster.*

5. Obtaining CMI-PB RNASeq data

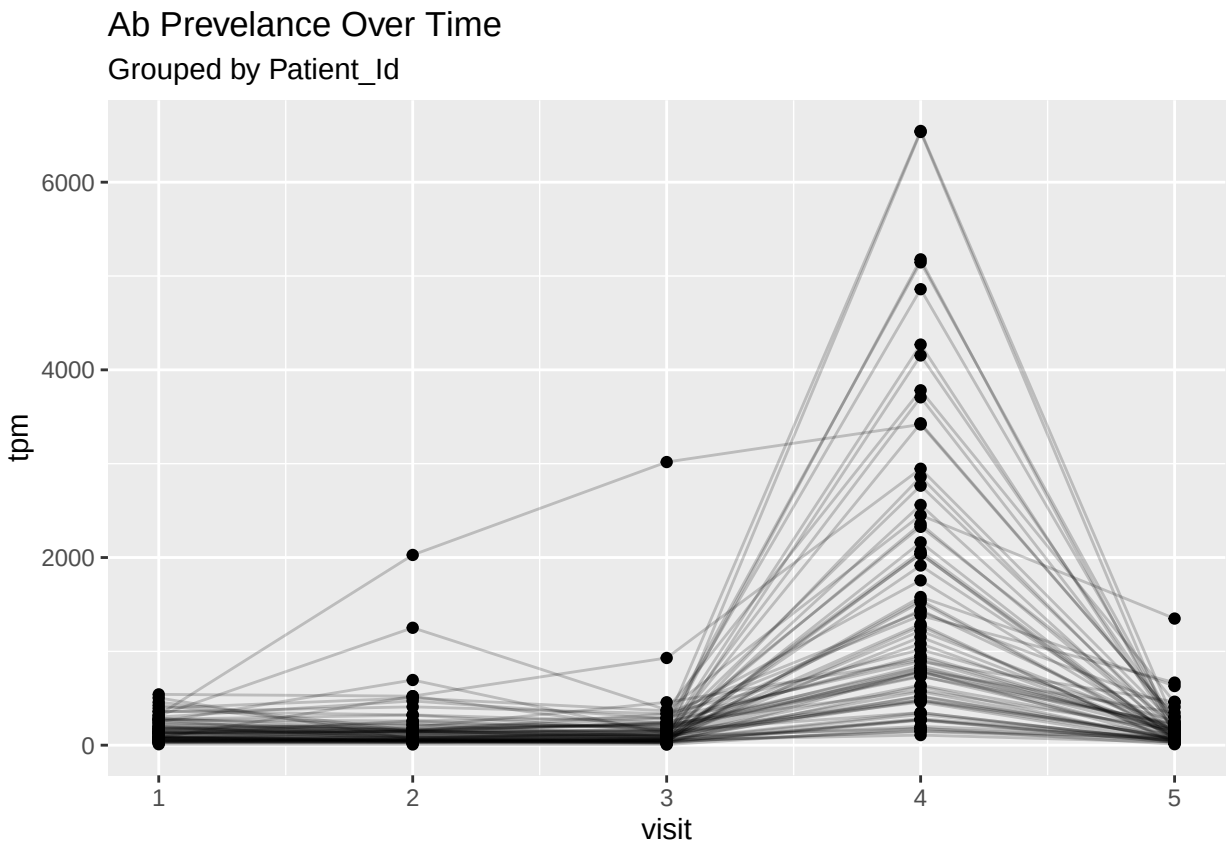
```
#Import RNA-Seq data
rna <- read_json("https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7",

#Join the RNA-Seq results to the patients table by 'specimen_id'
ssrna <- inner_join(rna, meta)
```

```
## Joining, by = "specimen_id"
```

Q18. Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit vs. tpm).

```
ggplot(ssrna) +
  aes(visit, tpm, group=subject_id) +
  geom_point() +
  geom_line(alpha=0.2) + labs(title="Ab Prevelance Over Time", subtitle="Grouped by Patient_Id")
```



Q19.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)? *The gene expression is maximum at visit 4*

Q20. Does this pattern in time match the trend of antibody titer data? If not, why not? *Yes, it matches. The highest RNA titer is at visit 4, and the highest antibody titer - although the RNA drops - is between visits 4 and 5. This makes sense because proteins stick around longer than mRNA.*