Investigating Pertussis Resurgence - Mini Project

Yash Garodia

08/03/2022

Background

Pertussis (more commonly known as whooping cough) is a highly contagious respiratory disease caused by the bacterium Bordetella pertussis. People of all ages can be infected leading to violent coughing fits followed by a high-pitched intake of breath that sounds like "whoop". Infants and toddlers have the highest risk for severe complications and death. Recent estimates from the WHO suggest that ~16 million cases and 200,000 infant deaths are due to pertussis annually 1.

1. Investigating pertussis cases by year

The United States Centers for Disease Control and Prevention (CDC) has been compiling reported pertussis case numbers since 1922 in their National Notifiable Diseases Surveillance System (NNDSS). We can view this data on the CDC website here: https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html

> 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(datapasta)
```

```
## Warning: package 'datapasta' was built under R version 4.0.2
```

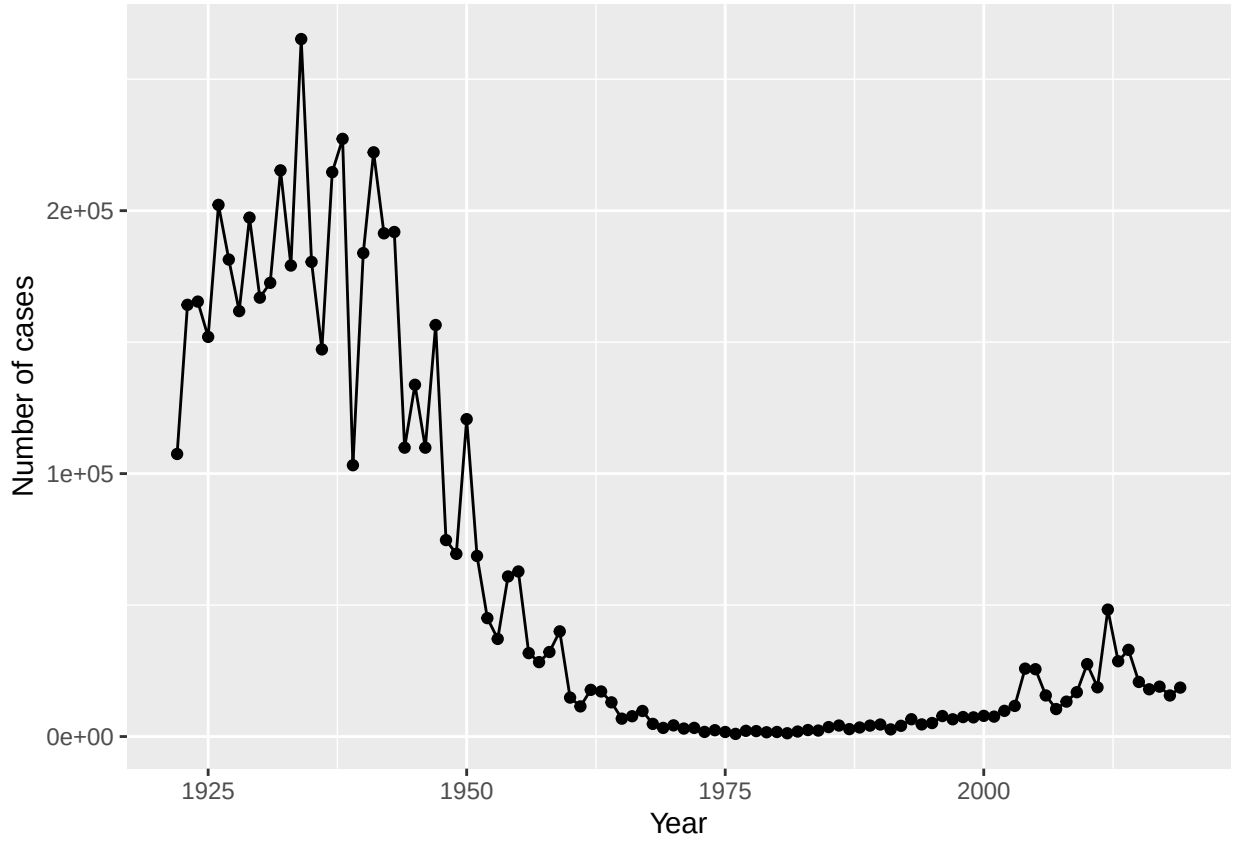
```
cdc <- data.frame(
                           Year = c(1922L,
                                    1923L,1924L,1925L,1926L,1927L,1928L,
                                    1929L,1930L,1931L,1932L,1933L,1934L,1935L,
                                    1936L,1937L,1938L,1939L,1940L,1941L,
                                    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)
)
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
ggplot(cdc) + aes(Year, No..Reported.Pertussis.Cases) + geom_point() + geom_line() + labs(x = "Year", y
```



2. A tale of two vaccines (wP & aP)

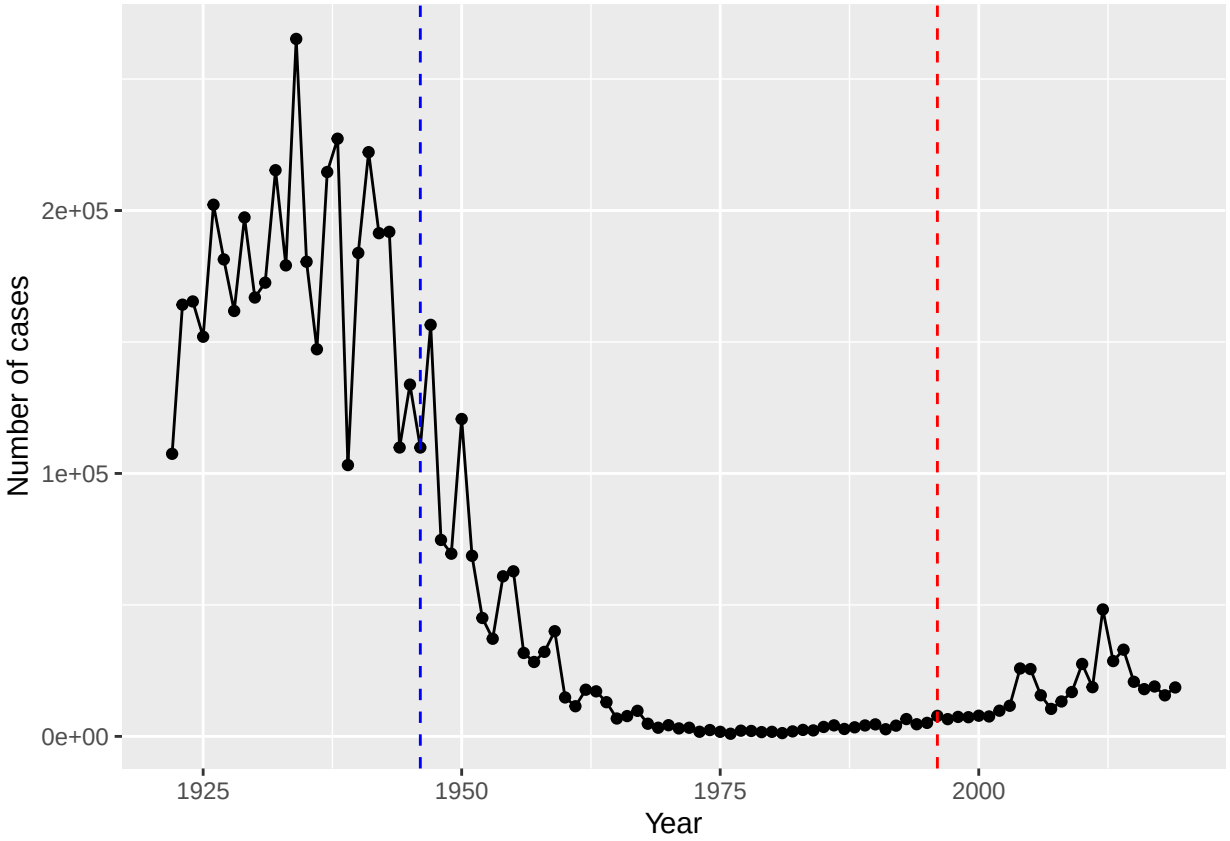
Two types of pertussis vaccines are currently available: whole-cell pertussis (wP) and acellular pertussis (aP). The first vaccines were composed of 'whole cell' (wP) inactivated bacteria. The latter aP vaccines use purified antigens of the bacteria (the most important pertussis components for our immune system). These aP vaccines were developed to have less side effects than the older wP vaccines and are now the only form administered in the United States

For motivated readers there is a nice historical account of the wP to aP vaccine switch in the US and elsewhere here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4975064/

Let's return to our CDC data plot and examine what happened after the switch to the acellular pertussis (aP) vaccination program.

> 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?

```
ggplot(cdc) + aes(Year, No..Reported.Pertussis.Cases) + geom_point() + geom_line() + labs(x = "Year", y
```



> Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend?

After the introduction of the aP vaccine, the number of cases observed started increasing. This could be due to the fact that the PCR based pertussis testing became more sensitive, so more number of cases were detected.

3. Exploring CMI-PB data

Why is this vaccine-preventable disease on the upswing? To answer this question we need to investigate the mechanisms underlying waning protection against pertussis. This requires evaluation of pertussis-specific immune responses over time in wP and aP vaccinated individuals.

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 returns JSON data

The CMI-PB API (like most APIs) sends responses in JSON format. Briefly, JSON data is formatted as a series of key-value pairs, where a particular word ("key") is associated with a particular value.

To read these types of files into R we will use the read_json() function from the jsonlite package. Note that if you want to do more advanced querys of APIs directly from R you will likely want to explore the more full featured rjson package. The big advantage of using jsonlite for our current purposes is that it can simplify JSON key-value pair arrays into R data frames without much additional effort on our part.

```
# Allows us to read, write and process JSON data
library(jsonlite)
```

```
## Warning: package 'jsonlite' was built under R version 4.0.2
```

Let's now read the main subject database table from the CMI-PB API. You can find out more about the content and format of this and other tables here: https://www.cmi-pb.org/blog/understand-data/

```
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
```

Q4. How may aP and wP infancy vaccinated subjects are in the dataset?

```
# We can use the table function for this.
table(subject$infancy_vac)
```

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

There are 47 ap vaccinated and 49 wP vaccinated subjects in the dataset.

Q5. How many Male and Female subjects/patients are in the dataset?

```
#This time we choose the biological_sex category.
table(subject$biological_sex)
```

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

There are 66 female and 30 male patients in the dataset.

Q6. What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc...)?

```
# In order to do this we use table on subject for both race and sex together.

table(subject$race, subject$biological_sex)
```

```
##
##                                         Female Male
##   American Indian/Alaska Native              0    1
##   Asian                                     18    9
##   Black or African American                  2    0
##   More Than One Race                         8    2
##   Native Hawaiian or Other Pacific Islander  1    1
##   Unknown or Not Reported                   10    4
##   White                                     27   13
```

7,8,9 are optional.

Joining multiple tables

Read the specimen and ab_titer tables into R and store the data as specimen and titer named data frames.

```
# Complete the API URLs...
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)
```

To know whether a given specimen_id comes from an aP or wP individual we need to link (a.k.a. "join" or merge) our specimen and subject data frames. The excellent dplyr package (that we have used previously) has a family of join() functions that can help us with this common task:

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:

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
##
## 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
```

```
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
```

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.

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

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

```
dim(abdata)
```

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

Q11. How many specimens (i.e. entries in abdata) do we have for each isotype?

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

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

We have 6698 specimens for IgE, 1413 for IgG, and 6141 for IgG2, IgG3, IgG4.

Q12. What do you notice about the number of visit 8 specimens compared to other visits?

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

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

The number of visit 8 specimens are lower than other visit specimens, and this could be because the study is ongoing.

#4. Examine IgG1 Ab titer levels

Now using our joined/merged/linked abdata dataset filter() for IgG1 isotype and exclude the small number of visit 8 entries.

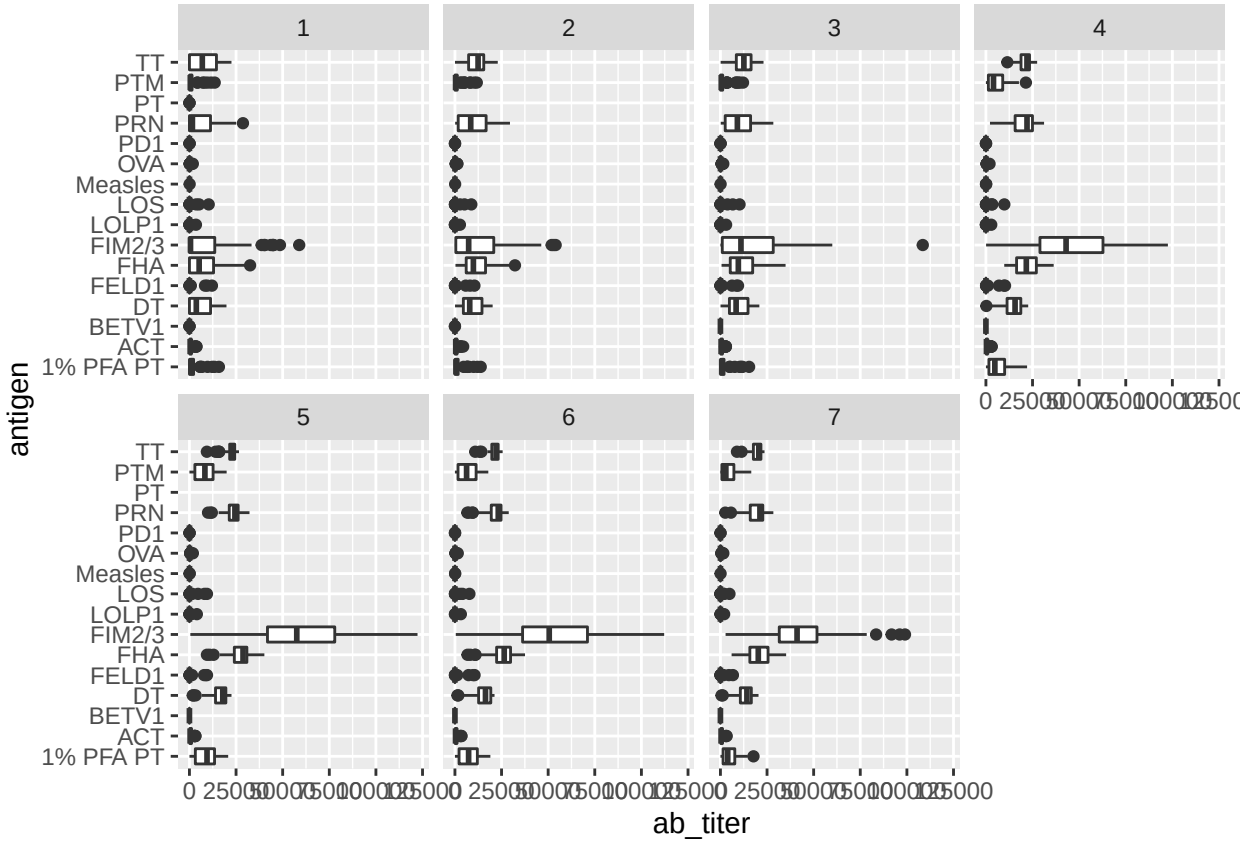
```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

```
##   specimen_id isotype is_antigen_specific antigen   ab_titer  unit
## 1           1    IgG1                TRUE     ACT 274.355068 IU/ML
## 2           1    IgG1                TRUE     LOS  10.974026 IU/ML
## 3           1    IgG1                TRUE   FELD1   1.448796 IU/ML
## 4           1    IgG1                TRUE   BETV1   0.100000 IU/ML
## 5           1    IgG1                TRUE   LOLP1   0.100000 IU/ML
## 6           1    IgG1                TRUE Measles  36.277417 IU/ML
##   lower_limit_of_detection subject_id actual_day_relative_to_boost
## 1                 3.848750          1                           -3
## 2                 4.357917          1                           -3
## 3                 2.699944          1                           -3
## 4                 1.734784          1                           -3
## 5                 2.550606          1                           -3
## 6                 4.438966          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
## 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
```

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)
```

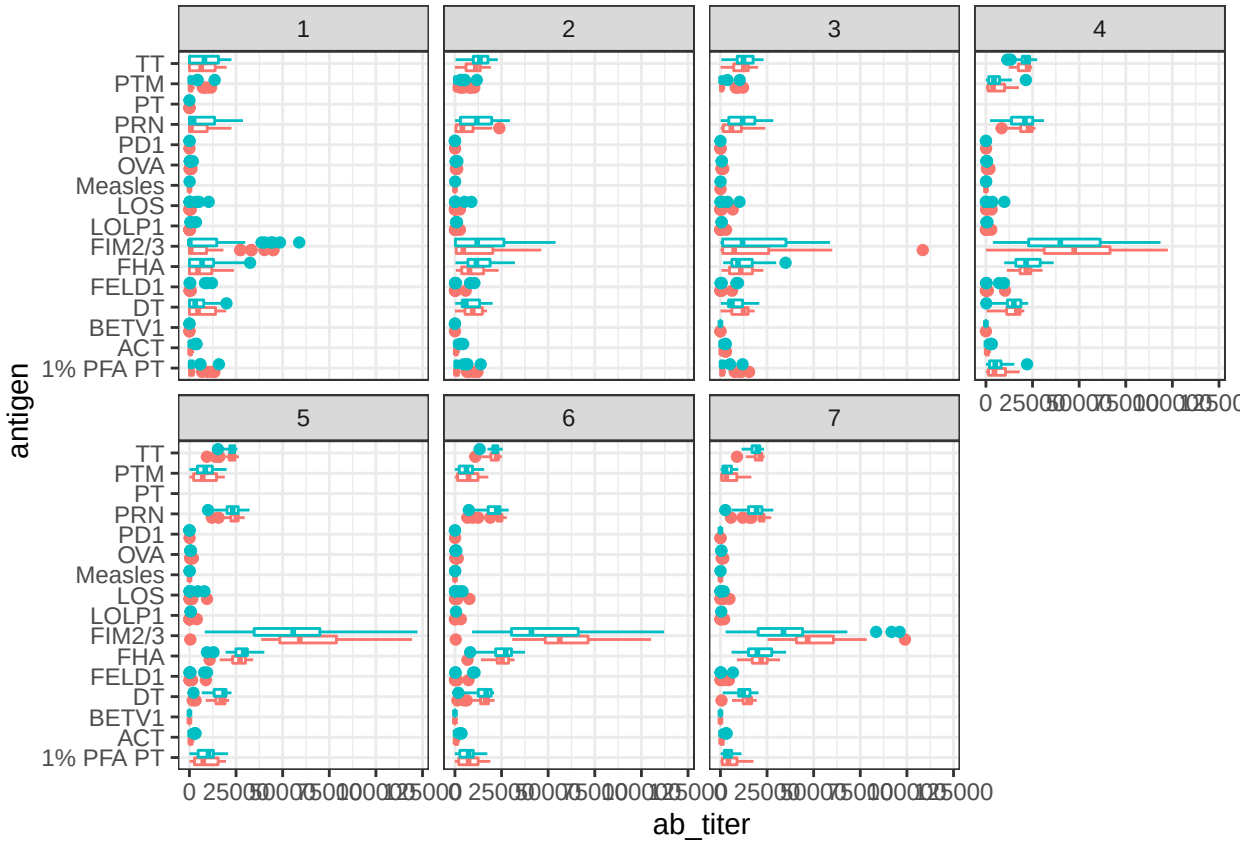


Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

FIM 2/3 and FHA increase in the levels of IgG1 antibody titers over time. This would be because both these antigens are in the antibody.

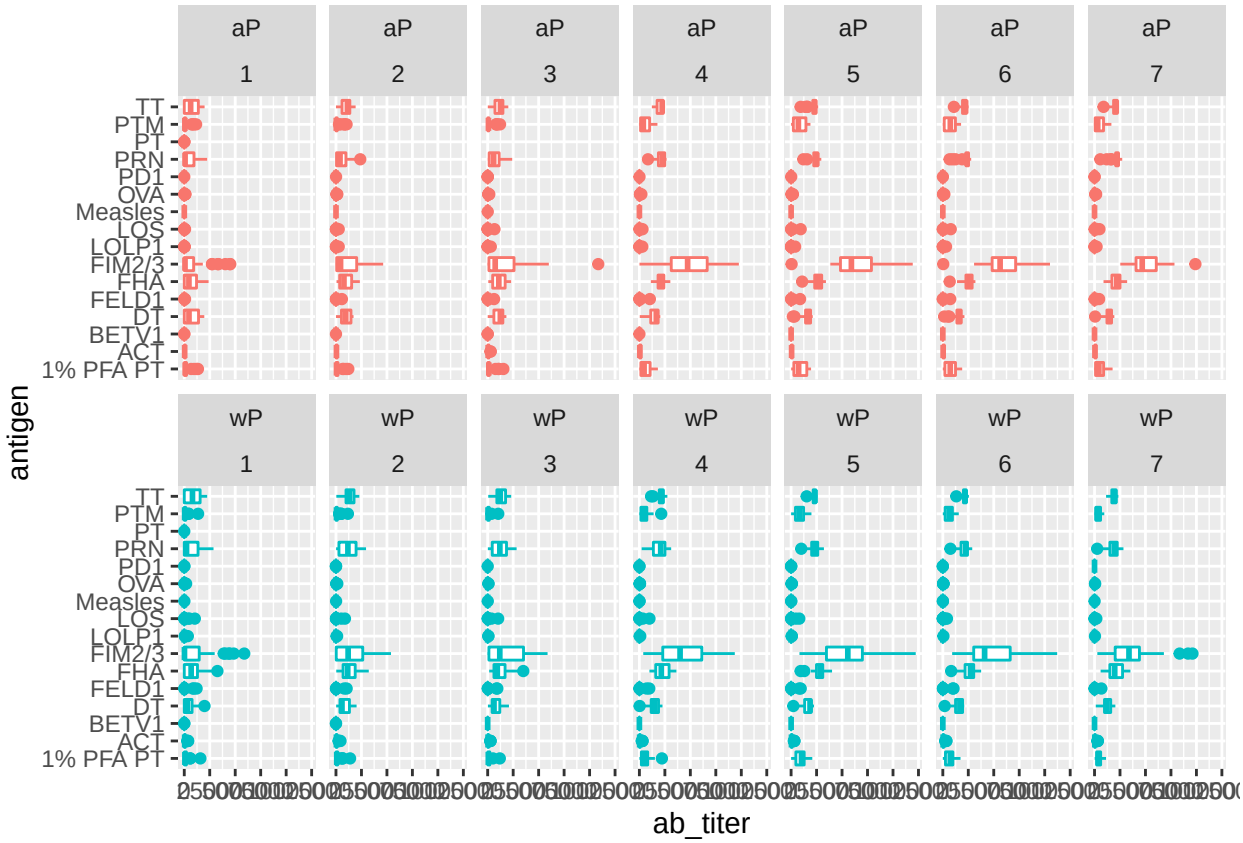
We can attempt to examine differences between wP and aP here by setting color and/or facet values of the plot to include infancy_vac status (see below). However these plots tend to be rather busy and thus hard to interpret easily.

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



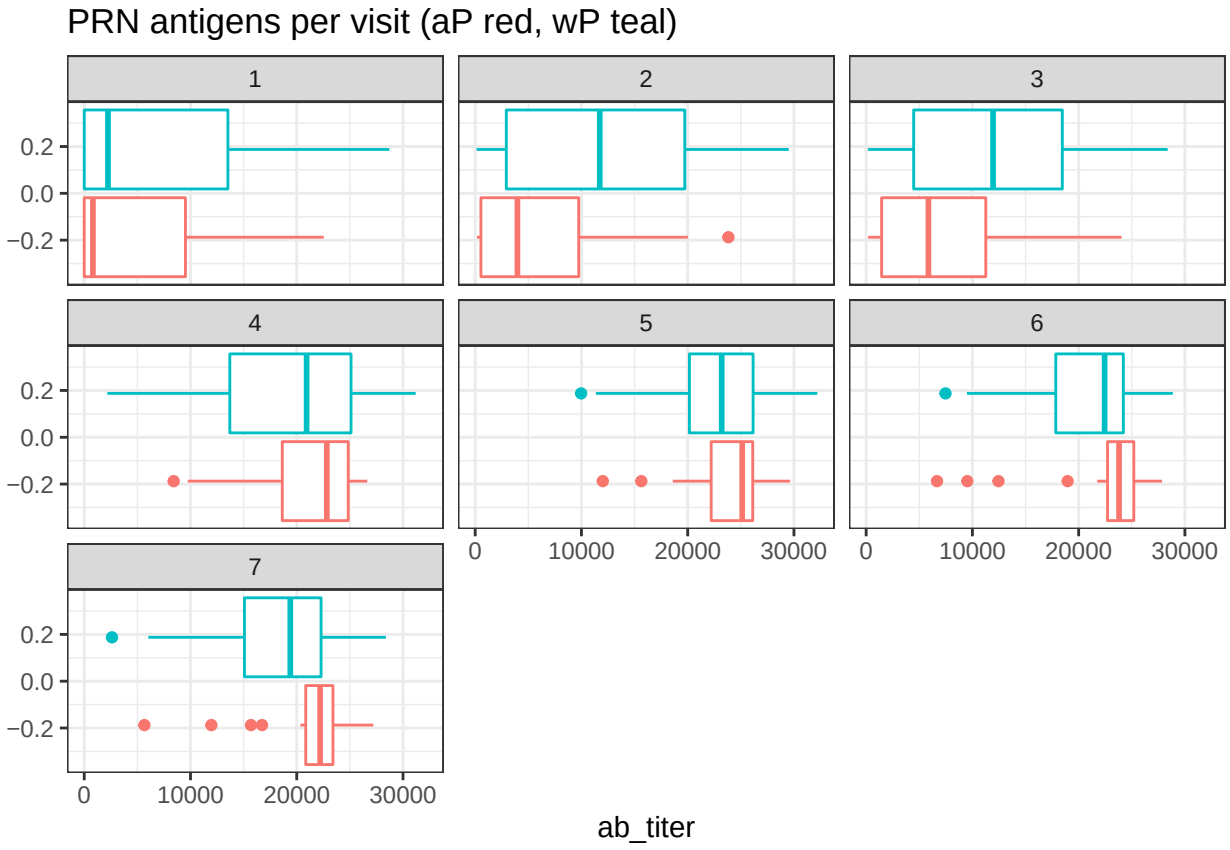
Another version of this plot adding infancy_vac to the faceting:

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

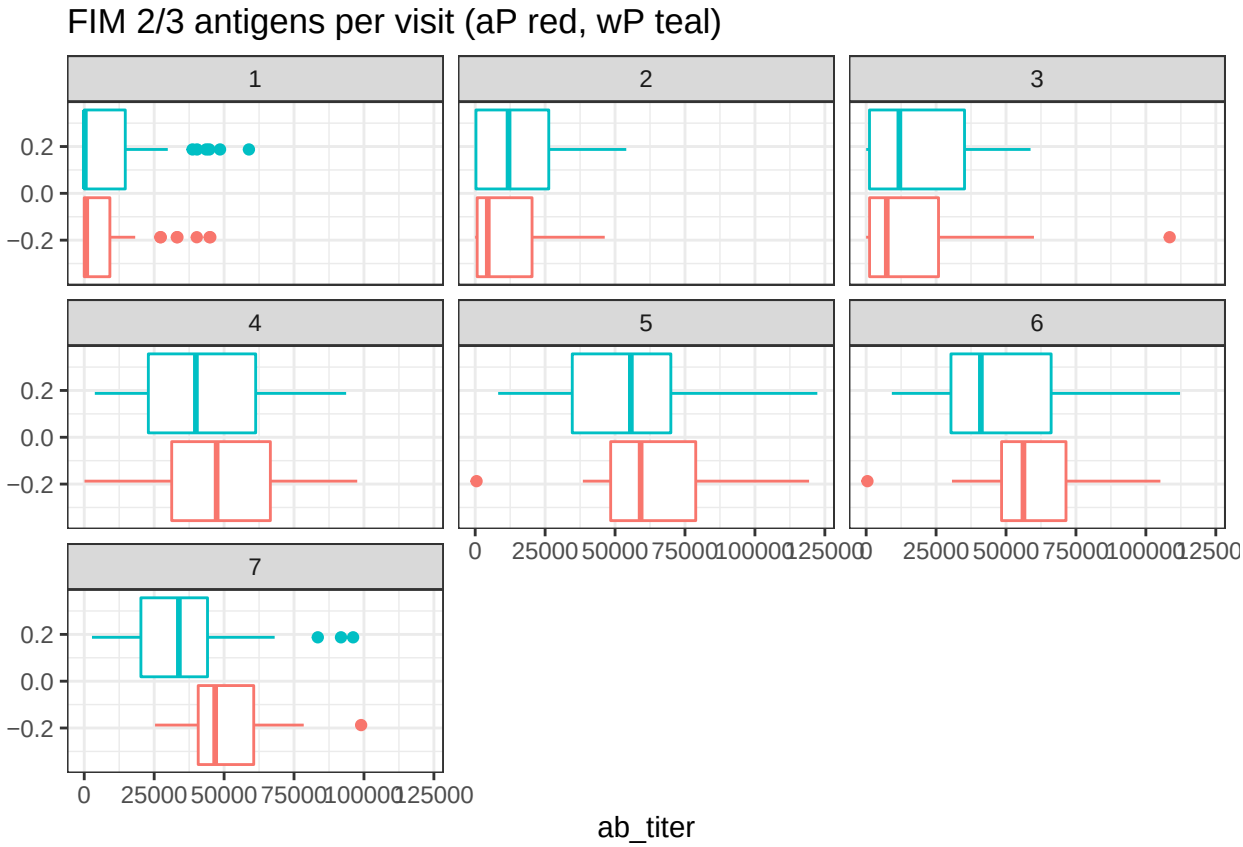


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like.

```
filter(ig1, antigen=="PRN") %>%
  ggplot() +
  aes(ab_titer, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() + labs (title = "PRN antigens per visit (aP red, wP teal)")
```




```
filter(ig1, antigen=="FIM2/3") %>%
  ggplot() +
  aes(ab_titer, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() + labs (title = "FIM 2/3 antigens per visit (aP red, wP teal)")
```



Q16. What do you notice about these two antigens time course and the FIM2/3 data in particular?

As time passes, both FIM 2/3 and PRN levels increase. FIM 2/3 peaks at visit 4, after which it seems to decline. PRN peaks at visit 5 as well, after which it seems to decline.

Q17. Do you see any clear difference in aP vs. wP responses?

wP responses are higher than aP responses.

5. Obtaining CMI-PB RNASeq data

For RNA-Seq data the API query mechanism quickly hits the web browser interface limit for file size. We will present alternative download mechanisms for larger CMI-PB datasets in the next section. However, we can still do "targeted" RNA-Seq querys via the web accessible API.

For example we can obtain RNA-Seq results for a specific ENSEMBLE gene identifier or multiple identifiers combined with the & character:

```
# For example use the following URL
#https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7
```

The link above is for the key gene involved in expressing any IgG1 antibody, namely the IGHG1 gene. Let's read available RNA-Seq data for this gene into R and investigate the time course of it's gene expression values.

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"

rna <- read_json(url, simplifyVector = TRUE)
```

To facilitate further analysis we need to "join" the rna expression data with our metadata meta, which is itself a join of sample and specimen data. This will allow us to look at this genes TPM expression values over aP/wP status and at different visits (i.e. times):

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

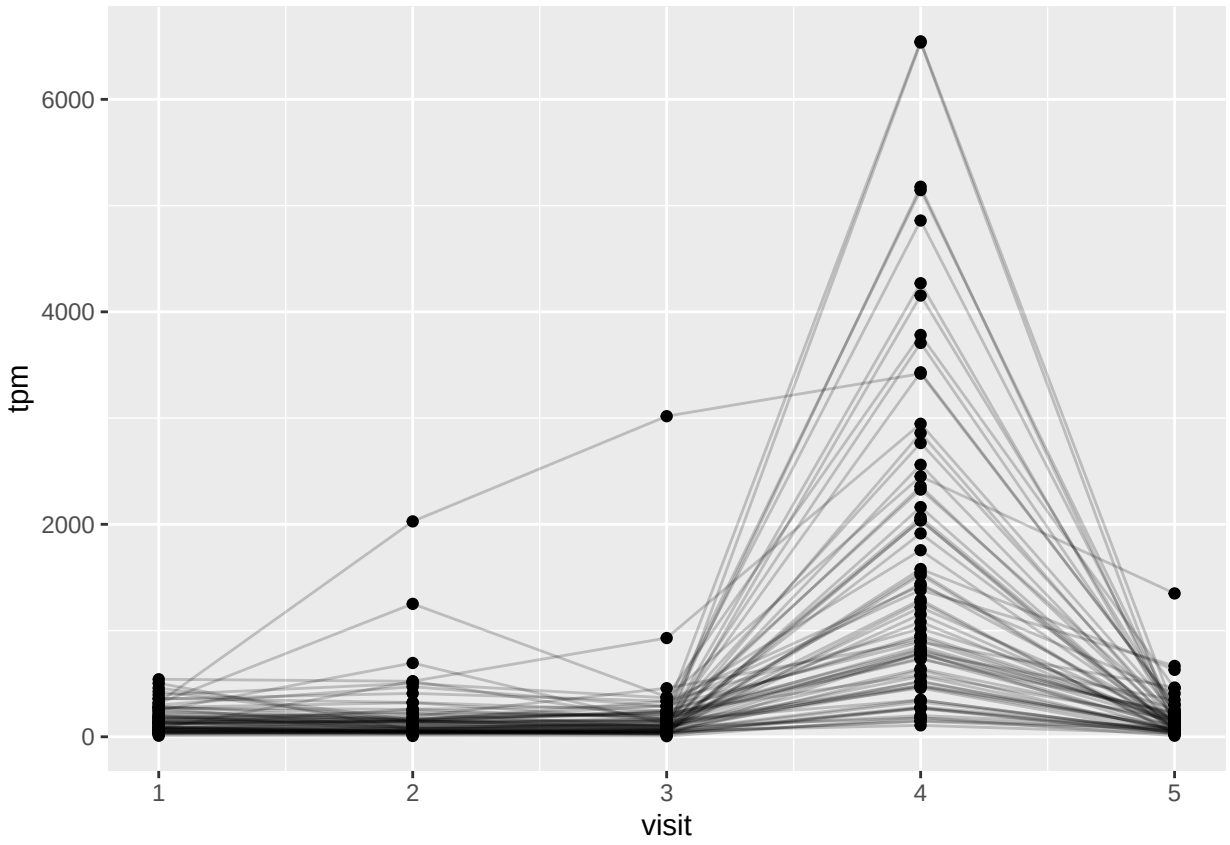
```
## Joining, by = "subject_id"
```

```
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)
```



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

Its maximum expression is on visit 4

Q20. Does this pattern in time match the trend of antibody titer data? If not, why not?

Antibody titer data suggests maximum expression on visit 5, they match because you need to express the gene for visit 4 in order for the antigen to be expressed by visit