Classlab 15: Mini Project: Investigating Pertussis Resurgence

Nathaniel Nono (PID: A16782656)

Background

- Systems vaccinology - Trying to understand how the immune system works and its relationship with vaccines
- Pertussis - Whopping cough; High contagious lung infection caused by the bacteria Bordetella pertussis
  - 16 million cases and 200,000 associated infant deaths annually
  - Can infect people of all ages but is most severe and life threatening for infants under a year old
  - Transmission occurs primarily through bacteria laden respiratory droplets
- Pertussis develops in three main phases
  - Catarrhal phase - Early symptoms; Runny nose cough, highly contagious
    - Antibiotics used as treatment
  - Paroxysmal phase - Severe symptoms; Paroxysms, whopping sound, exhaision
    - Antibiotics can help but it more so prevents the spread
  - Convalescent phase - Recovery phase
- Different vacinnes
  - Whole cell vaccines (wP) vaccine
  - Acellular (aP) vaccine
    - FHA - Adhesion proteins
- History of vaccines
  - 1578: First epidemic record
  - 1679: The Name “Pertussis” First Appears
  - 1900: Discovery of Bordetella pertussis → First observed that it was a bacteria
  - 1906: Causative Bacteria Isolated

  - 1942: First DPT Vaccine causing a decline in cases in the next 30 years
  - 1970s - 1980s: Antivax movements and massive lawsuits causing a rise in the disease
  - 1986: Nation childhood vaccine injury act
  - 1992: aP Vaccine Approved in the U.S.
  - 2010 - present: Pertussis outbreak in infants
  - CMI-PB Project: A new systems vaccinology project is launced that combines systems biology and genomics to provide a more holistic picture of protective pertussis-specific immune mechanisms. The project provides the scientific community with comprehensive, high-quality, and freely accessible resources related to Pertussis booster vaccination.

Pertussis, aka whopping chouch, is a high infection disease cause by the bactera *B. Pertussis*

The CDC tracks pertussis cases numbers per year. Let's have a closer look at this data: CDC data

1. Investigating pertussis cases by year

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.

Trouble, Data is in a pdf format :(

So, we will use the `datapasta` R pacakge to "scrape" this data into R:

- Install package in console: `install.packages('datapasta')`
- Copy table into clipboard
- Go into addins at the top -> `Paste as data.frame`

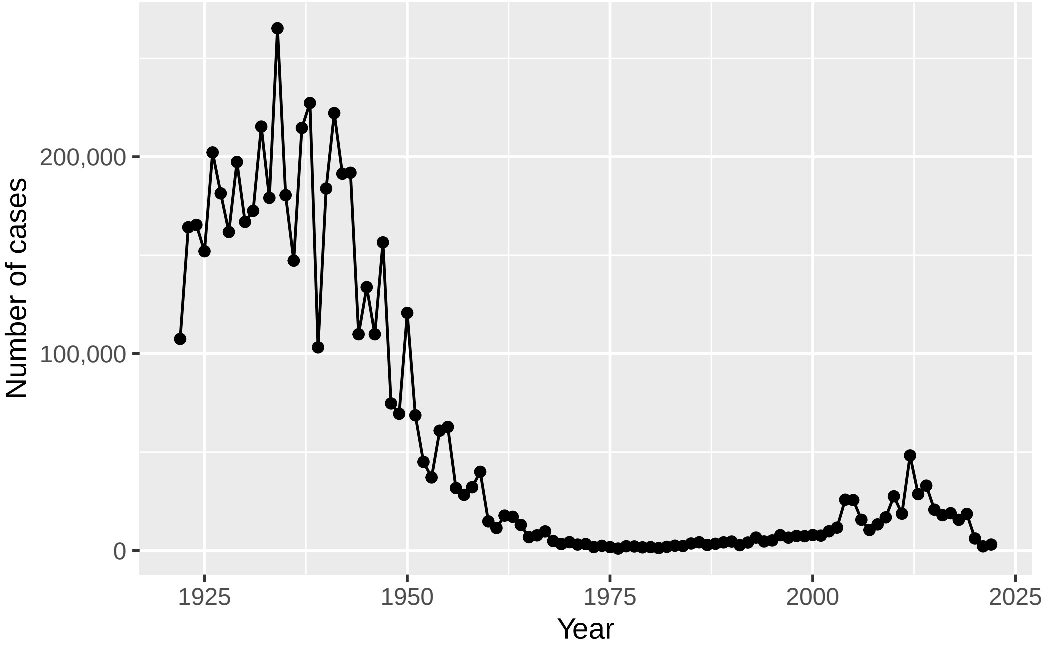
```
# Getting the dataframe from a pdf
cdc <- data.frame(
  year = c(
    1922L, 1923L, 1924L, 1925L,
    1926L, 1927L, 1928L, 1929L, 1930L, 1931L,
    1932L, 1933L, 1934L, 1935L, 1936L,
    1937L, 1938L, 1939L, 1940L, 1941L, 1942L,
    1943L, 1944L, 1945L, 1946L, 1947L,
    1948L, 1949L, 1950L, 1951L, 1952L,
    1953L, 1954L, 1955L, 1956L, 1957L, 1958L,
```

```r
    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, 2020L, 2021L, 2022L
  ),
  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,
    6124, 2116, 3044
  )
)
```

```r
# Call the ggplot2 package
library(ggplot2)
library(scales)
```

```
# Building the plot
baseplot <- ggplot(cdc) +
  aes(x = year,
      y = cases) +
  geom_point() +
  geom_line() +
  labs(x = "Year",
       y = "Number of cases") +
  scale_y_continuous(labels = comma) # No longer scientific notation

baseplot
```



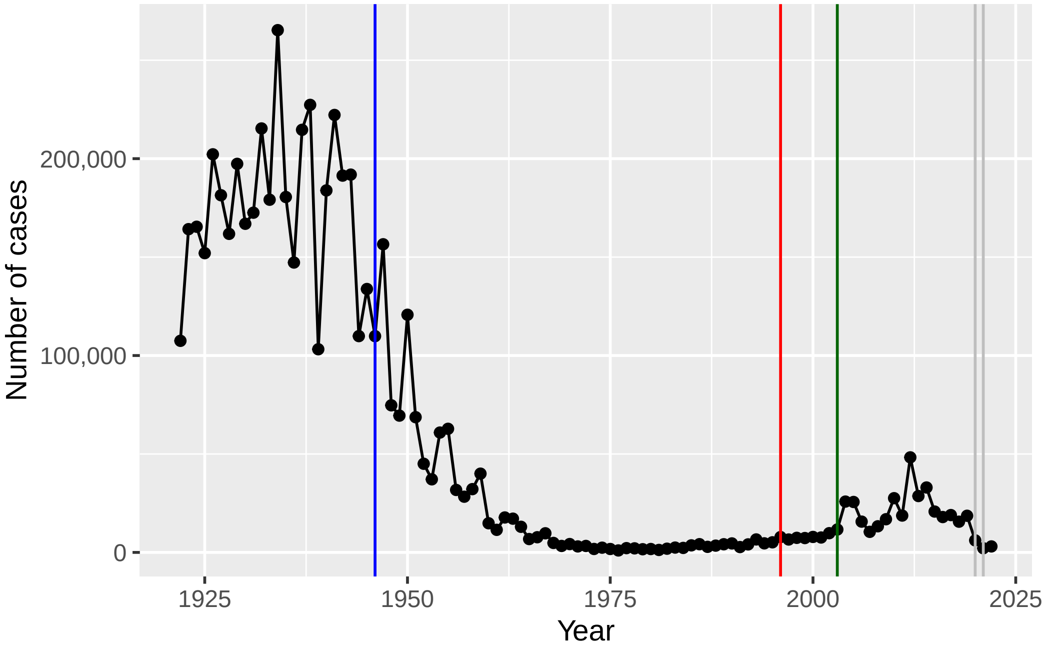
2. A tale of two vaccines (wP & aP)

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?

1) First whole-cell vaccine (wP) roll out in 1940
2) Switch to acellular vaccine (aP) in 1996
3) Covid in 2020-2021

```r
# Landmark plot
lm_plot <- baseplot +
  geom_vline(xintercept = 1946, # wC vaccine with everything
             col = 'blue') +
  geom_vline(xintercept = 1996, # aP vaccine with "essential components"
             col = 'red') +
  geom_vline(xintercept = 2003, # Start of the big increase
             col = 'darkgreen') +
  geom_vline(xintercept = c(2020,2021), # Covid-19 lockdowns
             col = 'grey')

lm_plot
```



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

We went from ~200,000 cases prewP vaccine to ~1,000 cases in 1976. However after the introduction of the aP vaccine we see a slight shift upwards in the number of cases after ~10 years with a big increase in 2004. This could be due to the sparked controversy of vaccines and an uprise in antivax movements, bacterial evolution due to an increase amount of antibiotic use, or the aP vaccine is not as effective

(not as long lasting). And we see the last one as there is ~10 year lag from a roll out to increasing case numbers

Key Question: Why does the aP vaccine induced immunity wane faster than that of the wP vaccine?

3. Exploring CMI-PB data

The CMI-PB (Computational Models of Immunity Pertussis Boost) makes avaialble lots of data about the immune reposne to Pertussis booster vaccination

Critically, it tracks wP and aP individuals over time to see how their immune response changes

The new and ongoing CMI-PB project aims to provide the scientific community with this very information: CMI-PB

We have datasets from a total of seven assays, each accompanied by its corresponding metadata. All experimental data and metadata are stored and managed in a relational database management system (RDBMS): Data Composition

To study the long-term effects of priming between the acellular-pertussis (aP) vs. whole-cellular pertussis (wP) vaccines, we have recruited individuals born prior to 1995 and those born after: Study Outline

---

Trouble again… Data is in a JSON format :(

So, we will use the `jsonlite` R pacakge to allow us to read, write and process JSON data

- Install package in console: `install.packages('jsonlite')`
- Call the package
- Use the function `read_json()` with the url in the parenthesis with quotes

```r
# Call package
library(jsonlite)

# Read subject table
subject <- read_json('https://www.cmi-pb.org/api/v5/subject',
                     simplifyVector = TRUE)
```

```
# Take a look of the table
head(subject)
```

```
  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
4          4          wP           Male Not Hispanic or Latino Asian
5          5          wP           Male Not Hispanic or Latino Asian
6          6          wP         Female Not Hispanic or Latino White
  year_of_birth date_of_boost      dataset
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
4    1988-01-01    2016-08-29 2020_dataset
5    1991-01-01    2016-08-29 2020_dataset
6    1988-01-01    2016-10-10 2020_dataset
```

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

Approach 1:

```
# aP vaccinated
sum(subject$infancy_vac == 'aP')
```

```
[1] 87
```

```
# aP vaccinated
sum(subject$infancy_vac == 'wP')
```

```
[1] 85
```

Approach 2:

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

```
aP wP
87 85
```

There are 87 aP vaccinated and 85 wP vaccinated

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

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

```

Female   Male
   112     60
```

There are 112 females and 60 males so the data is not really too representative of the entire population but we'll continue

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

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

```

                                            Female Male
  American Indian/Alaska Native                  0    1
  Asian                                         32   12
  Black or African American                      2    3
  More Than One Race                            15    4
  Native Hawaiian or Other Pacific Islander      1    1
  Unknown or Not Reported                       14    7
  White                                         48   32
```

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:

Subject data

```
head(subject)
```

```
  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
4          4          wP           Male Not Hispanic or Latino Asian
5          5          wP           Male Not Hispanic or Latino Asian
6          6          wP         Female Not Hispanic or Latino White
  year_of_birth date_of_boost      dataset
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
4    1988-01-01    2016-08-29 2020_dataset
5    1991-01-01    2016-08-29 2020_dataset
6    1988-01-01    2016-10-10 2020_dataset
```

Specimen data

```
specimens <- read_json('https://www.cmi-pb.org/api/v5/specimen',
                     simplifyVector = TRUE)
```

```
head(specimens)
```

```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           2          1                            1
3           3          1                            3
4           4          1                            7
5           5          1                           11
6           6          1                           32
  planned_day_relative_to_boost specimen_type visit
```

```
1                             0         Blood     1
2                             1         Blood     2
3                             3         Blood     3
4                             7         Blood     4
5                            14         Blood     5
6                            30         Blood     6
```

Noticed a similarity in the specimen and subject datasets. Can merge these two tables to make a new meta data

```
# Call the dplyr package
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
```

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

```
Joining with `by = join_by(subject_id)`
```

```
head(meta)
```

```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           2          1                            1
3           3          1                            3
4           4          1                            7
5           5          1                           11
6           6          1                           32
  planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
1                             0         Blood     1          wP         Female
```

```
2                             1         Blood     2          wP         Female
3                             3         Blood     3          wP         Female
4                             7         Blood     4          wP         Female
5                            14         Blood     5          wP         Female
6                            30         Blood     6          wP         Female
               ethnicity  race year_of_birth date_of_boost      dataset
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.

Titer data

```r
abdata <- read_json('http://cmi-pb.org/api/v5/plasma_ab_titer',
                    simplifyVector = TRUE)
```

```r
head(abdata)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgE               FALSE   Total 1110.21154       2.493425
2           1     IgE               FALSE   Total 2708.91616       2.493425
3           1     IgG                TRUE      PT   68.56614       3.736992
4           1     IgG                TRUE     PRN  332.12718       2.602350
5           1     IgG                TRUE     FHA 1887.12263      34.050956
6           1     IgE                TRUE     ACT    0.10000       1.000000
   unit lower_limit_of_detection
1 UG/ML                 2.096133
2 IU/ML                29.170000
3 IU/ML                 0.530000
4 IU/ML                 6.205949
5 IU/ML                 4.679535
6 IU/ML                 2.816431
```

One more join to do of `meta` and `abdata` to associate all the metadata about the individual and their race, biological sex, and inficnancy vaccination status together with Antibody levels…

```r
ab <- inner_join(abdata, meta)
```

```
Joining with `by = join_by(specimen_id)`
```

```r
head(ab)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgE               FALSE   Total 1110.21154       2.493425
2           1     IgE               FALSE   Total 2708.91616       2.493425
3           1     IgG                TRUE      PT   68.56614       3.736992
4           1     IgG                TRUE     PRN  332.12718       2.602350
5           1     IgG                TRUE     FHA 1887.12263      34.050956
6           1     IgE                TRUE     ACT    0.10000       1.000000
   unit lower_limit_of_detection subject_id actual_day_relative_to_boost
1 UG/ML                 2.096133          1                           -3
2 IU/ML                29.170000          1                           -3
3 IU/ML                 0.530000          1                           -3
4 IU/ML                 6.205949          1                           -3
5 IU/ML                 4.679535          1                           -3
6 IU/ML                 2.816431          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      dataset
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
```

Q. How many Ab measurments do we have?

```r
nrow(ab)
```

```
[1] 52576
```

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

```
table(ab$isotype)
```

```

  IgE   IgG  IgG1  IgG2  IgG3  IgG4 
 6698  5389 10117 10124 10124 10124 
```

> Q. How many antigens?

```
table(ab$antigen)
```

```

    ACT   BETV1      DT   FELD1     FHA  FIM2/3   LOLP1     LOS Measles     OVA 
   1970    1970    4978    1970    5372    4978    1970    1970    1970    4978 
    PD1     PRN      PT     PTM   Total      TT 
   1970    5372    5372    1970     788    4978 
```

Let's focus in on IgG - One of the main antibody types responsive to bacteria or virial infectoons

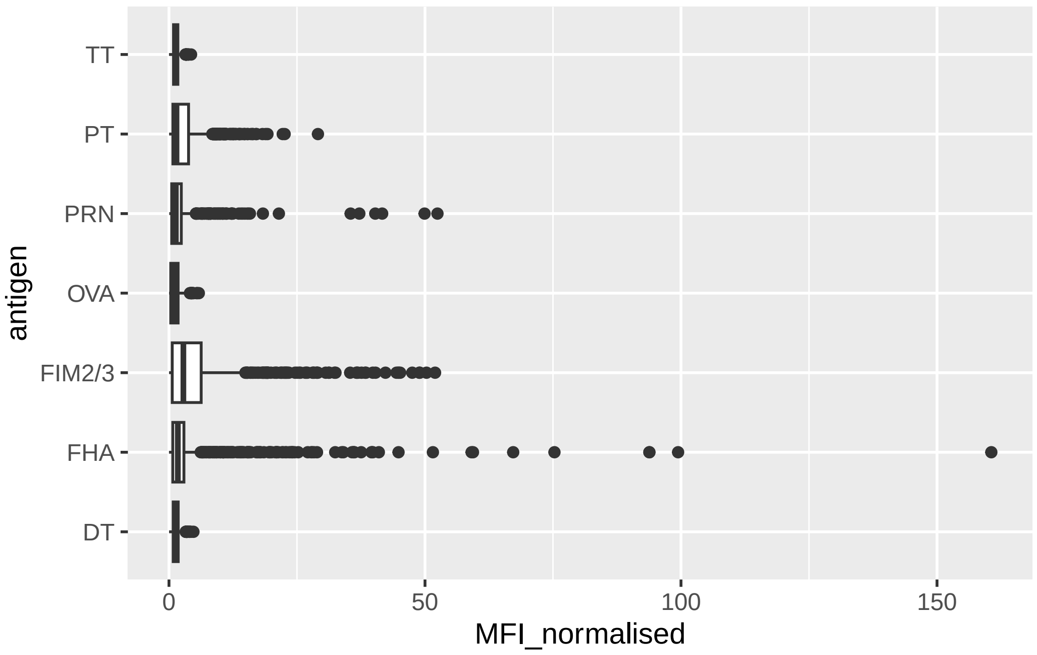
```
igg <- filter(ab, isotype == 'IgG')
head(igg)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgG                TRUE      PT   68.56614       3.736992
2           1     IgG                TRUE     PRN  332.12718       2.602350
3           1     IgG                TRUE     FHA 1887.12263      34.050956
4          19     IgG                TRUE      PT   20.11607       1.096366
5          19     IgG                TRUE     PRN  976.67419       7.652635
6          19     IgG                TRUE     FHA   60.76626       1.096457
   unit lower_limit_of_detection subject_id actual_day_relative_to_boost
1 IU/ML                 0.530000          1                           -3
2 IU/ML                 6.205949          1                           -3
3 IU/ML                 4.679535          1                           -3
4 IU/ML                 0.530000          3                           -3
5 IU/ML                 6.205949          3                           -3
6 IU/ML                 4.679535          3                           -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      dataset
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                Unknown White    1983-01-01    2016-10-10 2020_dataset
5                Unknown White    1983-01-01    2016-10-10 2020_dataset
6                Unknown White    1983-01-01    2016-10-10 2020_dataset
```

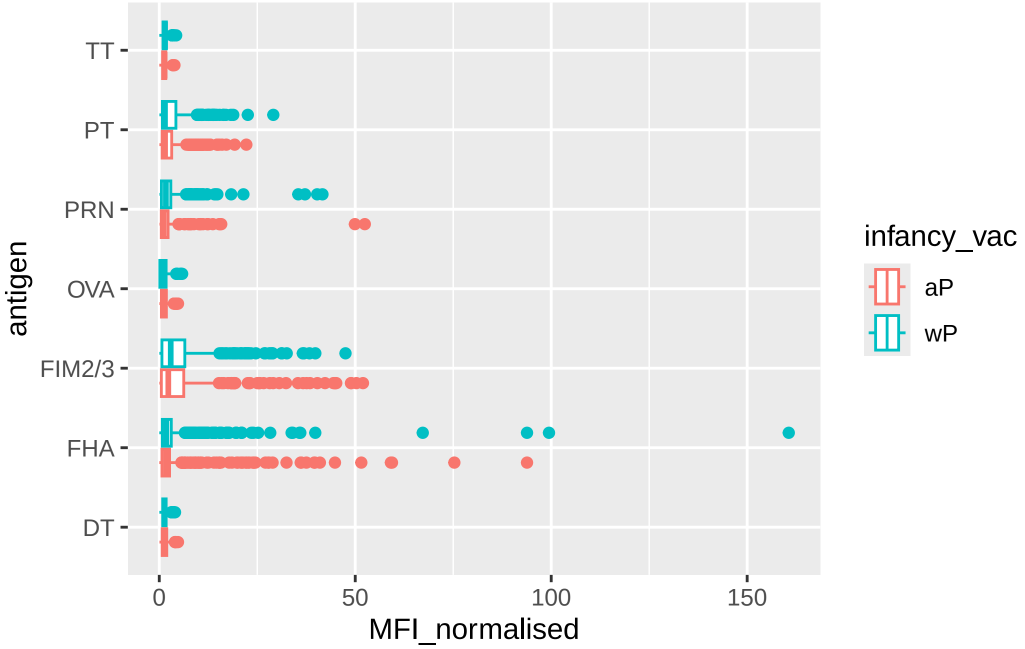
Make a first plot of Mean Florescence Intensity (MFI); measure of how muh is detected

```
ggplot(igg) +
  aes(MFI_normalised, antigen) +
  geom_boxplot()
```

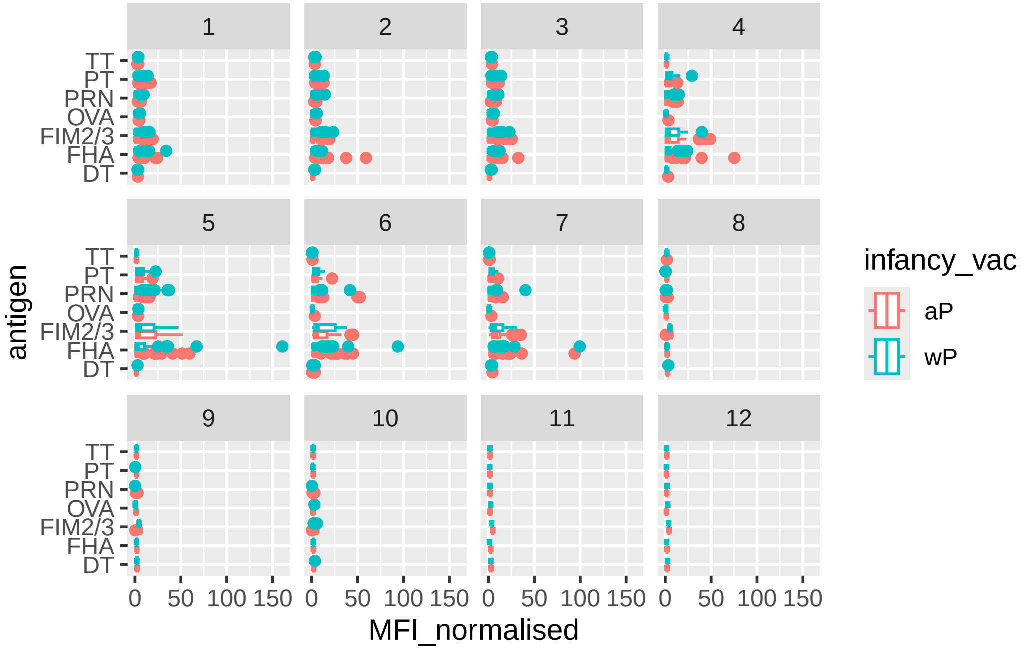


Let's color by aP/wP infancy_vac

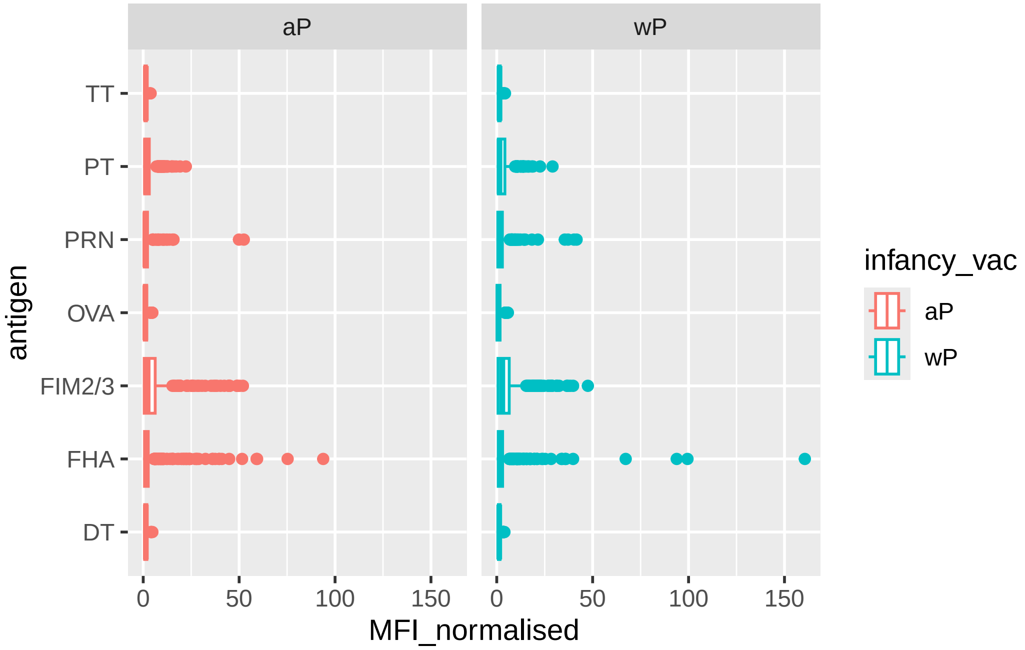
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col = infancy_vac) +
  geom_boxplot()
```



```
ggplot(igg) +
  aes(MFI_normalised, antigen, col = infancy_vac) +
  geom_boxplot() +
  facet_wrap(~visit) # Faceting by visit
```



```
ggplot(igg) +
  aes(MFI_normalised, antigen, col = infancy_vac) +
  geom_boxplot() +
  facet_wrap(~infancy_vac) # Faceting by vaccine
```



```
table(igg$visit)
```

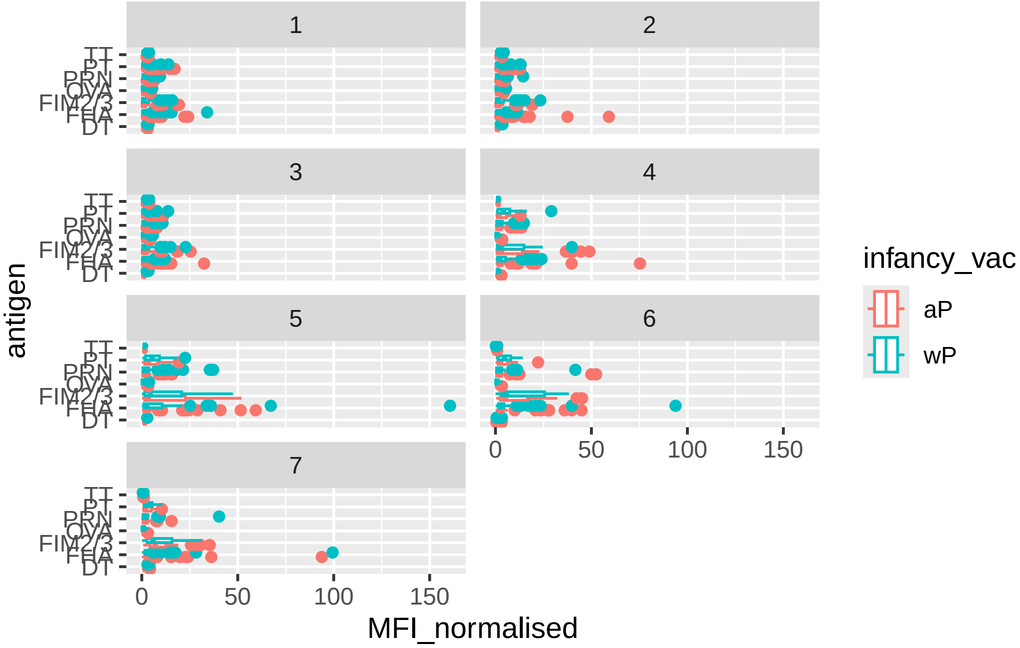
```
  1   2   3   4   5   6   7   8   9  10  11  12
902 902 930 559 559 540 525 150 147 133  21  21
```

There's a lot of visitation in the beginning but since the data is being constantly updated not all the patients have gone through all the visits. Let's focus solely on the first 7 visits and exclude visits 8-12 since they are not representative of the sample size

```
igg_7 <- filter(igg, visit %in% 1:7)
table(igg_7$visit)
```

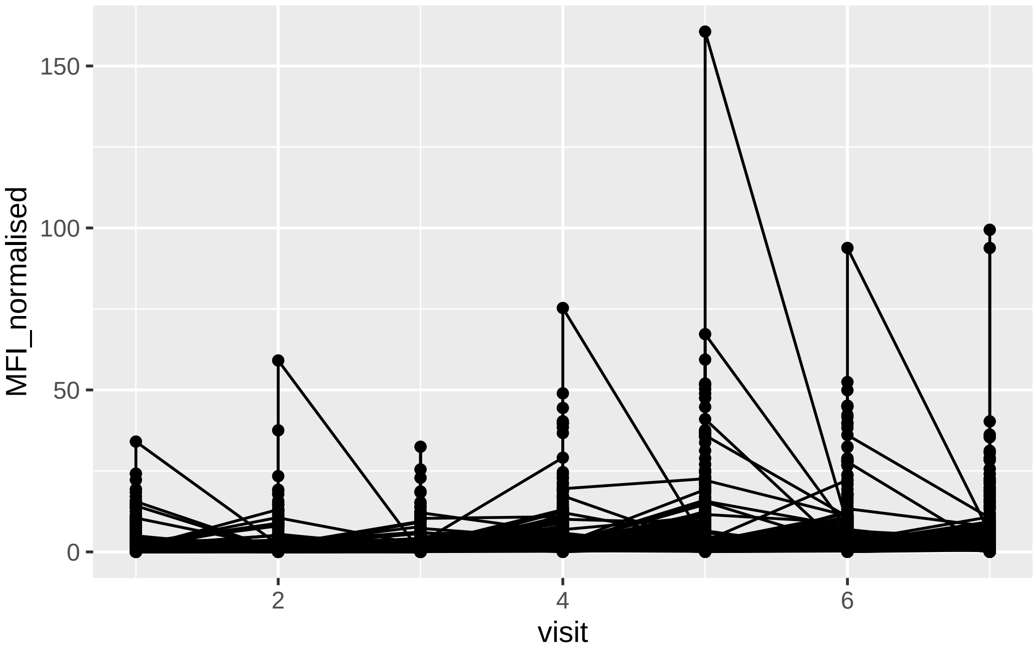
```
  1   2   3   4   5   6   7
902 902 930 559 559 540 525
```

```
ggplot(igg_7) +
  aes(MFI_normalised, antigen, col = infancy_vac) +
  geom_boxplot() +
  facet_wrap(~visit, ncol = 2) # Faceting by visit
```



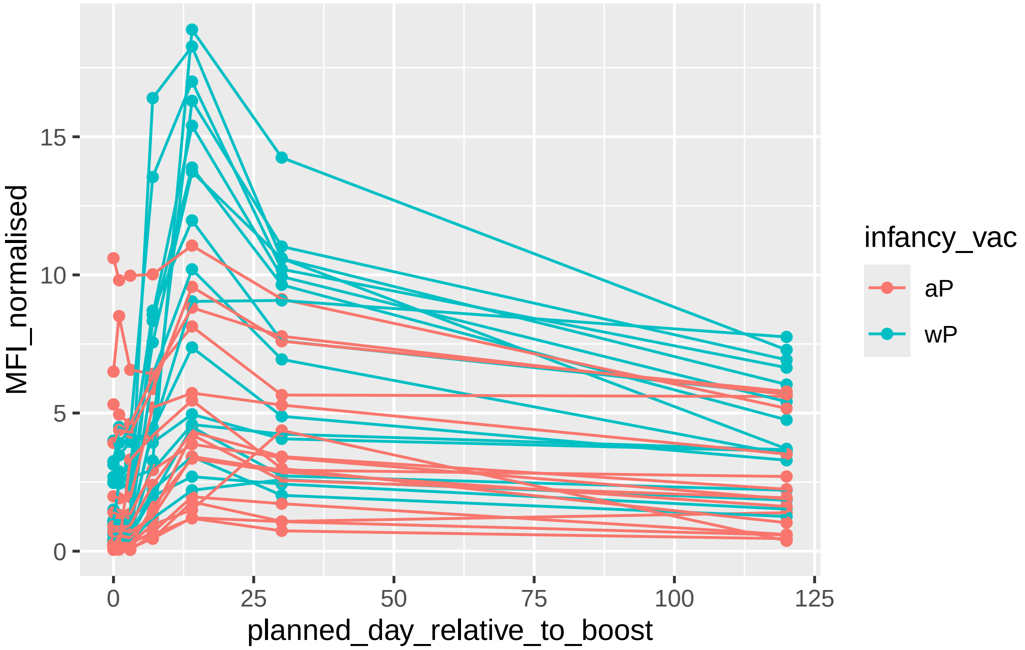
Let's try a different plot. First focus on one antgen, start with PT (Pertussis toxin) and plot visit or time on the x-axis and MFI_normalized on the y-axis.

```
ggplot(igg_7) +
  aes(x = visit,
      y = MFI_normalised,
      group=subject_id) +
  geom_point() +
  geom_line()
```



```
abdata.21 <- ab %>% filter(dataset == "2021_dataset")

abdata.21 %>%
  filter(isotype == "IgG",  antigen == "PT") %>%
  ggplot() +
    aes(x=planned_day_relative_to_boost,
        y=MFI_normalised,
        col=infancy_vac,
        group=subject_id) +
    geom_point() +
    geom_line()
```



Note: Let’s finish here today. We are behinning to see some interesting differences betrweeen aP and wP indiviudals. There is likely lots of other itneresting things to find in this dataset…

Not covered

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?

Q8. Determine the age of all individuals at time of boost?

Q9. With the help of a faceted boxplot or histogram (see below), do you think these two groups are significantly different?

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:

Q12. What are the different $dataset values in abdata and what do you notice about the number of rows for the most "recent" dataset?

4. Examine IgG Ab titer levels

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

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

Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the bacterium B. pertussis).

Q16. What do you notice about these two antigens time courses and the PT data in particular?

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

Q18. Does this trend look similar for the 2020 dataset?

5. Obtaining CMI-PB RNASeq data

Q19. Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit vs. tpm).

Q20.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

Q21. Does this pattern in time match the trend of antibody titer data? If not, why not?

6. Working with larger datasets [OPTIONAL]