Investigating Pertussis Resurgence

Nicole Alfonso, PID: A16429176

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)
library(ggplot2)
```

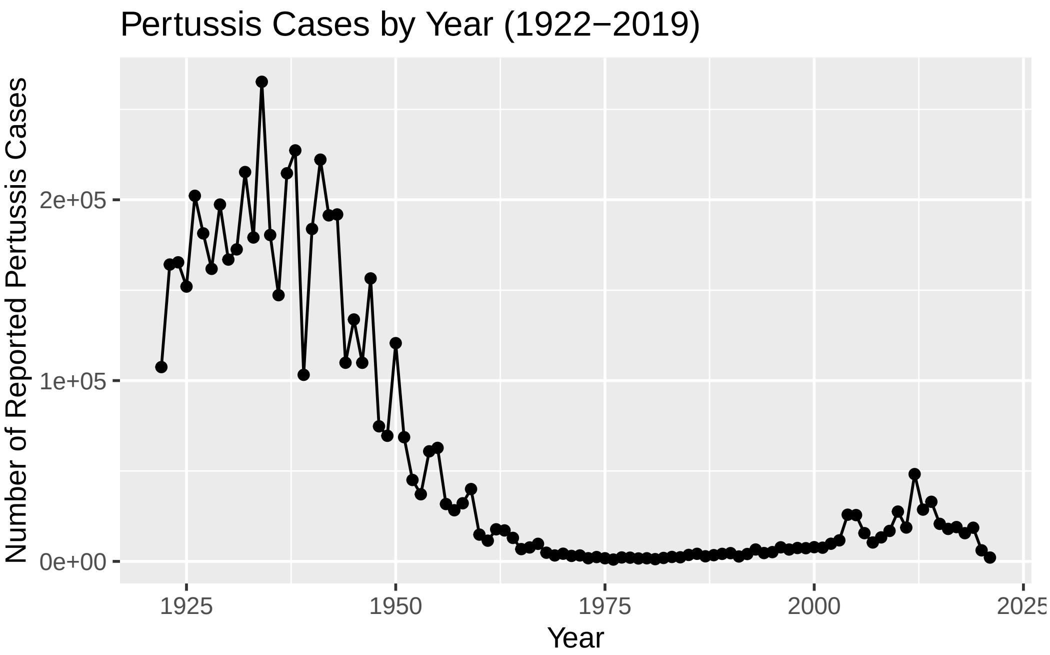
```
cdc <- data.frame(
                          Year = c(1922L,
                                   1923L,1924L,1925L,1926L,1927L,1928L,
                                   1929L,1930L,1931L,1932L,1933L,1934L,1935L,
                                   1936L,1937L,1938L,1939L,1940L,1941L,
                                   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,2020L,2021L),
  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)
)

cdc
```

```
    Year  Cases
1   1922 107473
2   1923 164191
3   1924 165418
4   1925 152003
5   1926 202210
6   1927 181411
7   1928 161799
8   1929 197371
9   1930 166914
10  1931 172559
11  1932 215343
12  1933 179135
13  1934 265269
14  1935 180518
15  1936 147237
16  1937 214652
17  1938 227319
18  1939 103188
19  1940 183866
```

20	1941	222202
21	1942	191383
22	1943	191890
23	1944	109873
24	1945	133792
25	1946	109860
26	1947	156517
27	1948	74715
28	1949	69479
29	1950	120718
30	1951	68687
31	1952	45030
32	1953	37129
33	1954	60886
34	1955	62786
35	1956	31732
36	1957	28295
37	1958	32148
38	1959	40005
39	1960	14809
40	1961	11468
41	1962	17749
42	1963	17135
43	1964	13005
44	1965	6799
45	1966	7717
46	1967	9718
47	1968	4810
48	1969	3285
49	1970	4249
50	1971	3036
51	1972	3287
52	1973	1759
53	1974	2402
54	1975	1738
55	1976	1010
56	1977	2177
57	1978	2063
58	1979	1623
59	1980	1730
60	1981	1248
61	1982	1895
62	1983	2463

```
63  1984   2276
64  1985   3589
65  1986   4195
66  1987   2823
67  1988   3450
68  1989   4157
69  1990   4570
70  1991   2719
71  1992   4083
72  1993   6586
73  1994   4617
74  1995   5137
75  1996   7796
76  1997   6564
77  1998   7405
78  1999   7298
79  2000   7867
80  2001   7580
81  2002   9771
82  2003  11647
83  2004  25827
84  2005  25616
85  2006  15632
86  2007  10454
87  2008  13278
88  2009  16858
89  2010  27550
90  2011  18719
91  2012  48277
92  2013  28639
93  2014  32971
94  2015  20762
95  2016  17972
96  2017  18975
97  2018  15609
98  2019  18617
99  2020   6124
100 2021   2116
```

```
ggplot(cdc) +
  aes(x = Year, y = Cases) +
  geom_point() +
  geom_line() +
  labs(title = "Pertussis Cases by Year (1922-2019)",
       x = "Year",
       y = "Number of Reported Pertussis Cases")
```



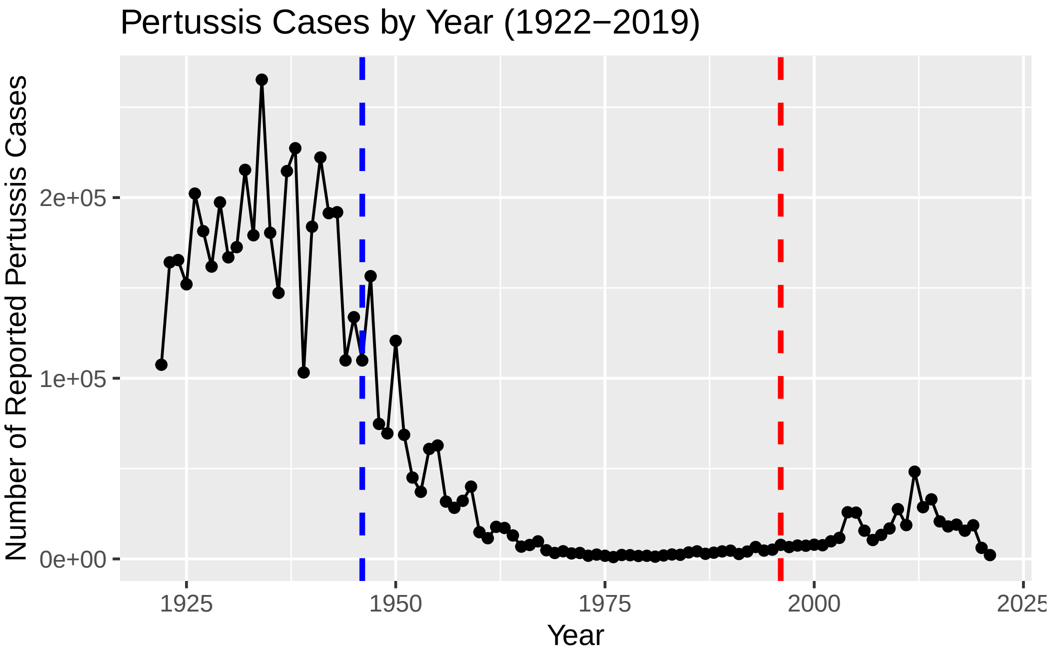
2. A tale of two vaccines (wP & aP)

Two types of pertussis vaccines have been developed: **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, see Figure 2). 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.

> 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. What do you notice?
>
> Looking at the generated plot, we observe the number of pertussis cases per year to decrease and plateu in the years following the introduction of both the wP and aP vaccines.

```r
ggplot(cdc, aes(x = Year, y = Cases)) +
  geom_line() +
  geom_point() +
  geom_vline(xintercept = 1946, linetype = "dashed", color = "blue", size = 1) +
  geom_vline(xintercept = 1996, linetype = "dashed", color = "red", size = 1) +
  labs(title = "Pertussis Cases by Year (1922-2019)",
       x = "Year",
       y = "Number of Reported Pertussis Cases")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



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 pertussis cases per year remain steady, followed by a spike in cases after about a decade. The rise in cases could allude to a new strain of the disease that has evolved, which the vaccine would not be effective against, or a drop in the number of vaccinations received by the general public. Additionally, it may be possible that the vaccine loses effectiveness after a few years, and thus, would require a second dose/booster shot.

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 CMI-PB API returns JSEON 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.

```
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      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
```

> Q4. How many aP and wP vaccinated subjects are in the data set?
>
> ap = 60, wp = 58

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

```

aP wP 
60 58 
```

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

female = 79, male = 39

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

```
Female   Male 
    79     39 
```

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

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

```
        
         American Indian/Alaska Native Asian Black or African American
  Female                             0    21                         2
  Male                               1    11                         0
        
         More Than One Race Native Hawaiian or Other Pacific Islander
  Female                  9                                         1
  Male                    2                                         1
        
         Unknown or Not Reported White
  Female                      11    35
  Male                         4    20
```

Side-Note: Working with dates

```
library(lubridate)
```

```
Attaching package: 'lubridate'

The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```r
# What is today's date
today()
```

```
[1] "2024-03-15"
```

```r
# How many days have passed since new year 2000
today() - ymd("2000-01-01")
```

```
Time difference of 8840 days
```

```r
# What is this in years?
time_length(today() - ymd("2000-01-01"), "years")
```

```
[1] 24.2026
```

> Q7. Using this approach, determine the average age of wP individuals, the average age of aP individuals, and are they significantly different?

```r
subject$age <- today() - ymd(subject$year_of_birth)
```

```r
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
```

```r
ap <- subject %>% filter(infancy_vac == "aP")

round( summary( time_length( ap$age, "years" ) ) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     21      26      26      26      27      30 
```

```
wp <- subject %>% filter(infancy_vac == "wP")
round( summary( time_length( wp$age, "years" ) ) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     28      31      36      37      39      56 
```

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

```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

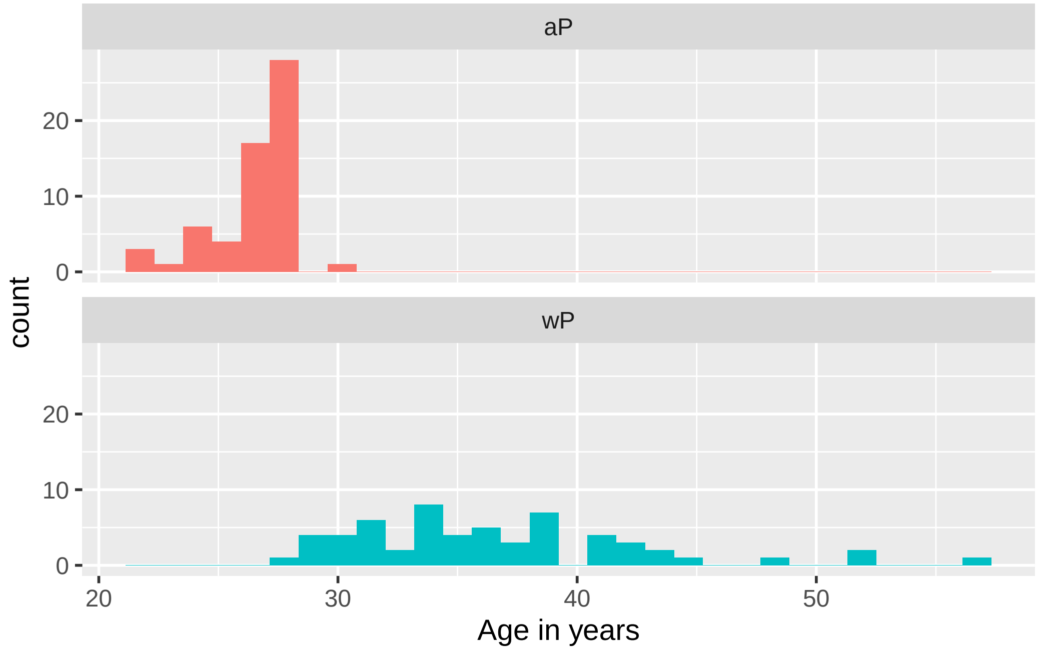
```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

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

Yes, these two groups are significantly different, as there is little to no overlap between the data, as well as the fact that they occupy different extremes of the age spectrum. The aP vaccine is administered much earlier than the wP vaccine.

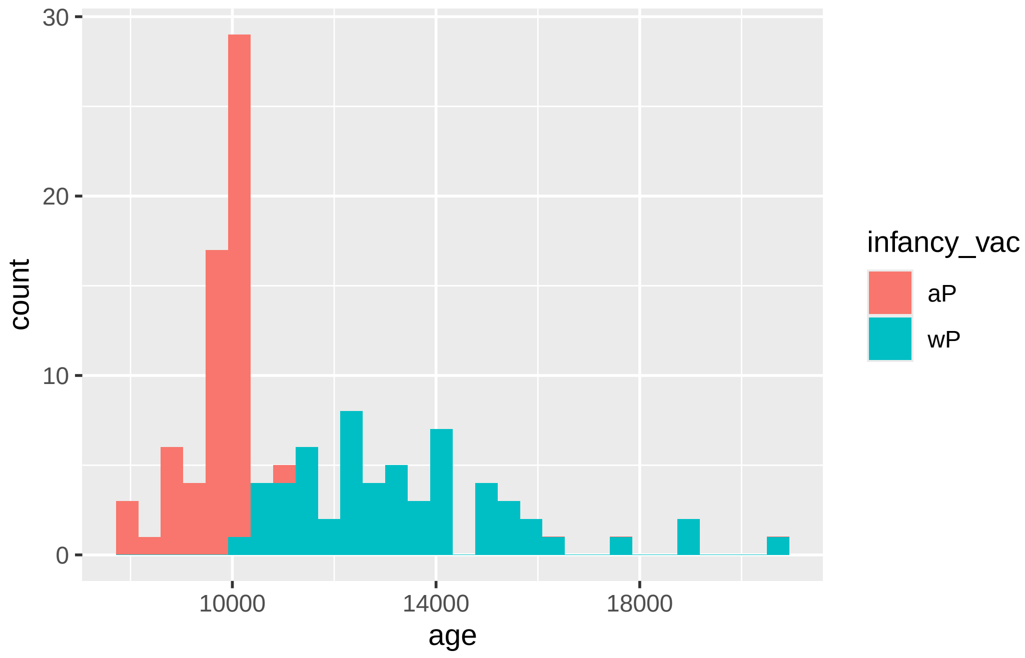
```
ggplot(subject) +
  aes(time_length(age, "year"),
      fill=as.factor(infancy_vac)) +
  geom_histogram(show.legend=FALSE) +
  facet_wrap(vars(infancy_vac), nrow=2) +
  xlab("Age in years")
```

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



```r
ggplot(subject) +
  aes(age, fill = infancy_vac) +
  geom_histogram()
```

```
Don't know how to automatically pick scale for object of type <difftime>.
Defaulting to continuous.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



To check the statistical significance: The difference is statistically significant.

```
x <- t.test(time_length( wp$age, "years" ),
       time_length( ap$age, "years" ))

x$p.value
```

```
[1] 6.813505e-19
```

Joining multiple tables

```
# Complete the API URLs...
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector = TRUE)
head(specimen)
```

```
  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
```

```
titer <- read_json("https://www.cmi-pb.org/api/v4/plasma_ab_titer", simplifyVector = TRUE)
head(titer)
```

```
  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
```

> 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

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

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

```
dim(meta)
```

```
[1] 939  14
```

```r
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
         age
1 13953 days
2 13953 days
3 13953 days
4 13953 days
5 13953 days
6 13953 days
```

> 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/femalte, etc.

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

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

```r
dim(abdata)
```

```
[1] 41775    21
```

> Q11. How many specimens do we have for each isotype?

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

```

 IgE  IgG IgG1 IgG2 IgG3 IgG4 
6698 3233 7961 7961 7961 7961 
```

> Q12. What are the different $dataset values in abdata and what do you notice about the number of rows for the most "recent" dataset?
>
> The data set for 2022 is much smaller than the 2020 dataset.

```
table(abdata$dataset)
```

```

2020_dataset 2021_dataset 2022_dataset 
       31520         8085         2170 
```

4. Examine IgG Ab titer levels

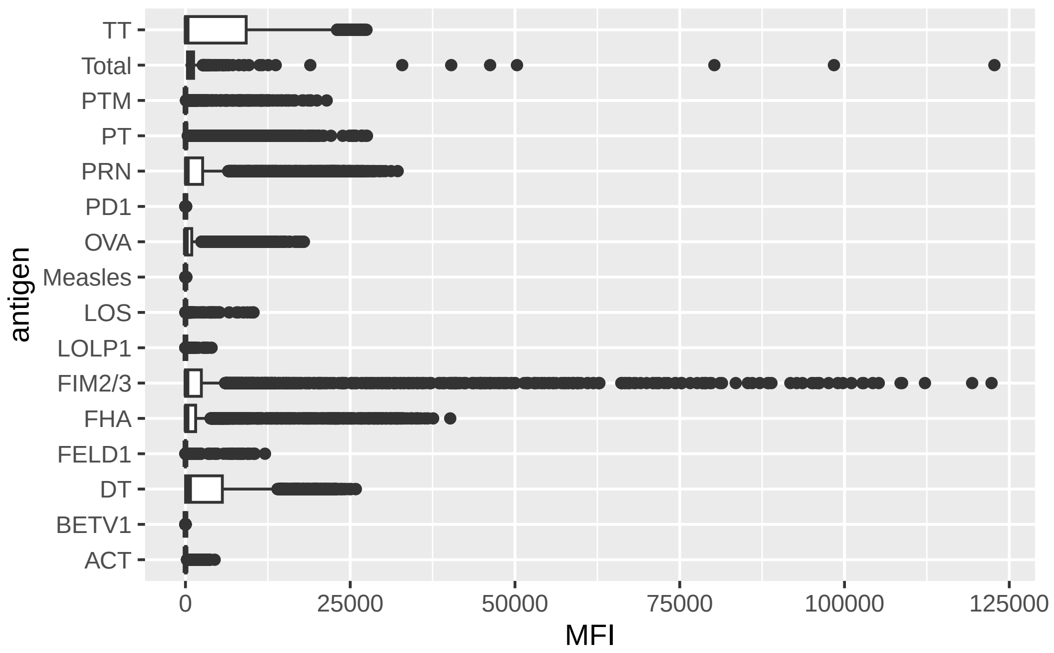
First exploratory plot

```
table(abdata$antigen)
```

```

    ACT   BETV1      DT   FELD1     FHA  FIM2/3   LOLP1     LOS Measles     OVA 
   1970    1970    3435    1970    3829    3435    1970    1970    1970    3435 
    PD1     PRN      PT     PTM   Total      TT 
   1970    3829    3829    1970     788    3435 
```

```
ggplot(abdata) +
    aes(MFI, antigen) +
    geom_boxplot()
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```

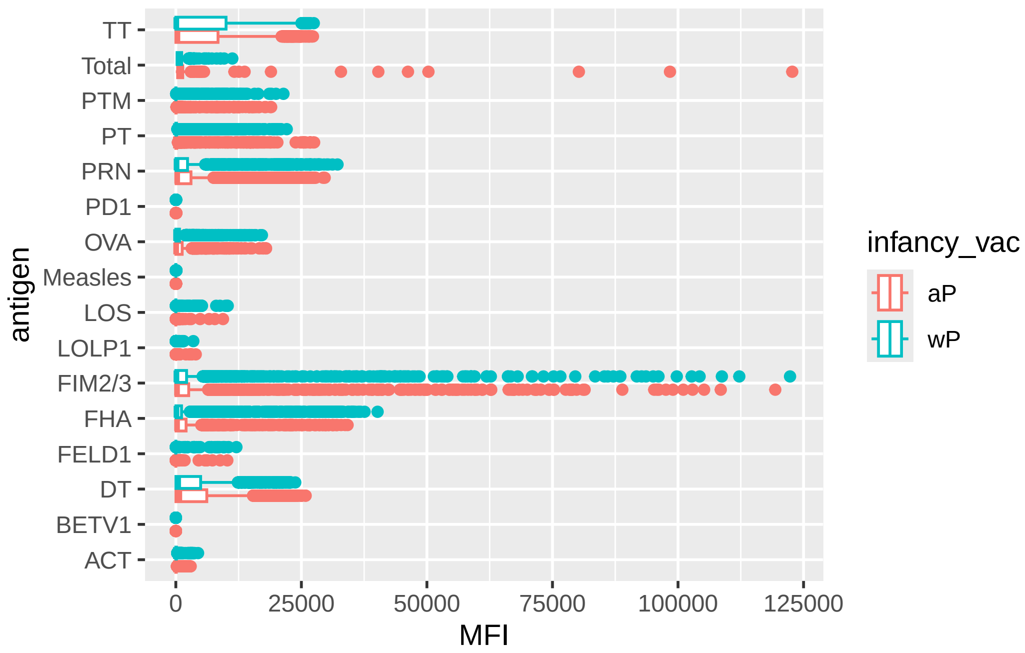


Certain antigens and not others are very variable in their detection levels – because the vaccine contained specific antibodies that led to variation in their detected levels

Can you facet or color by infancy_vac? Are there differences?

```
ggplot(abdata) +
   aes(MFI, antigen, col = infancy_vac) +
   geom_boxplot()
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```



```r
table(abdata$dataset)
```

```
2020_dataset 2021_dataset 2022_dataset 
       31520         8085         2170 
```

We will focus on different variables

2021 dataset

```r
 abdata.21 <- filter(abdata, dataset == "2021_dataset")
  table(abdata.21$dataset)
```

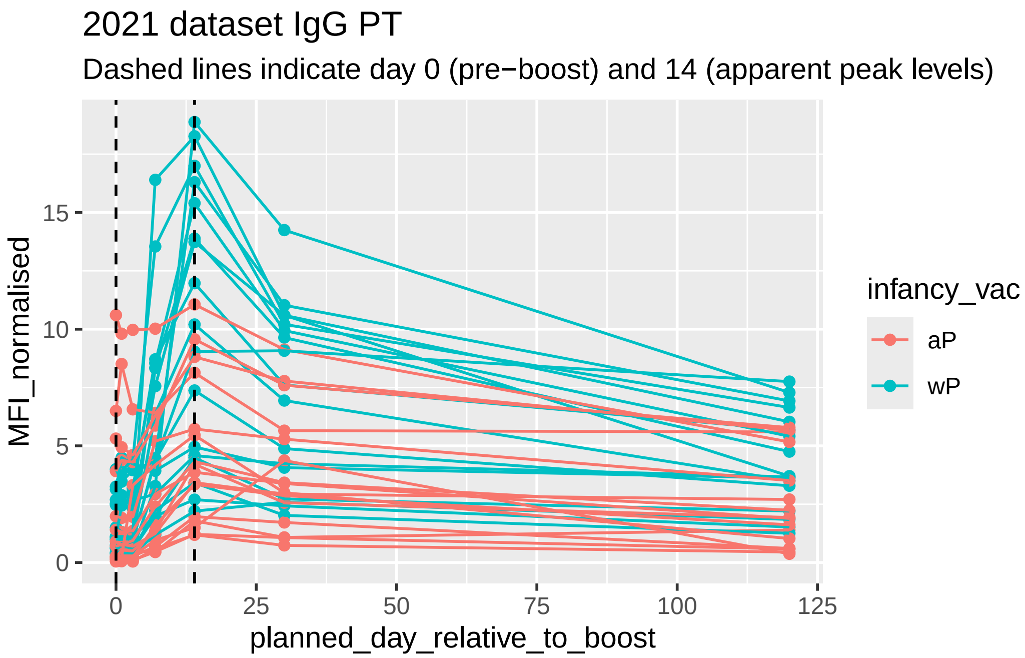
```
2021_dataset 
        8085 
```

PT antigen IgG levels

```r
pt.21 <- filter(abdata.21, isotype == "IgG", antigen == "PT")
```

We will compare days (time) to boost vs MFI levels

```
ggplot(pt.21) +
   aes(x=planned_day_relative_to_boost,
         y=MFI_normalised,
         col=infancy_vac,
         group = subject_id) +
   geom_point() +
   geom_line() +
   geom_vline(xintercept=0, linetype="dashed") +
   geom_vline(xintercept=14, linetype="dashed") +
   labs(title="2021 dataset IgG PT",
        subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)
```



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

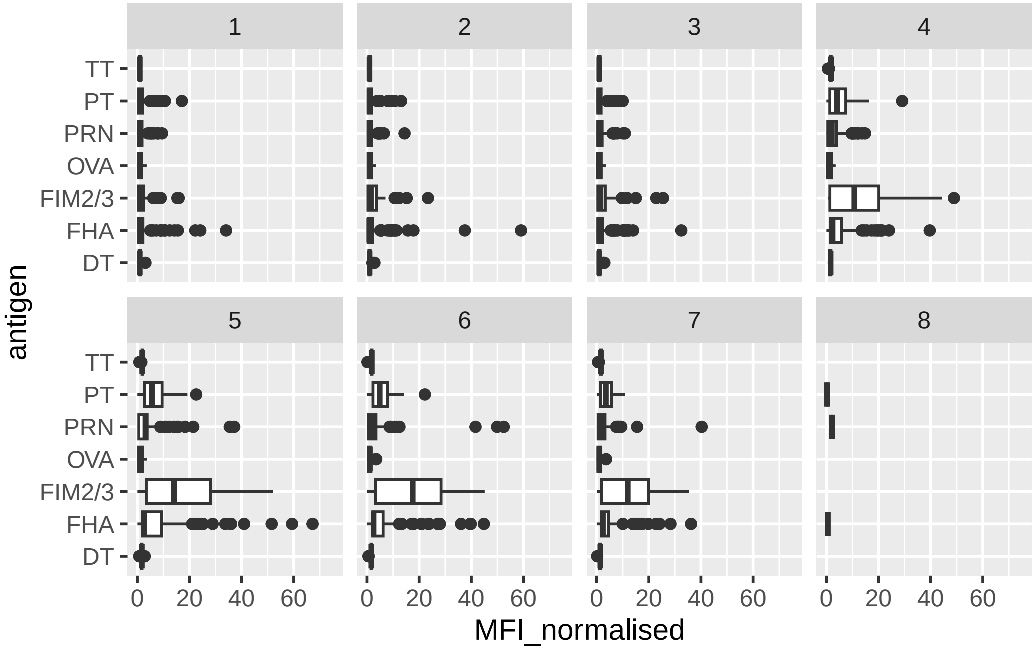
```
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

```
  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
         age
1 13953 days
2 13953 days
3 13953 days
4 15049 days
5 15049 days
6 15049 days
```

```r
ggplot(igg) +
   aes(MFI_normalised, antigen) +
   geom_boxplot() +
     xlim(0,75) +
   facet_wrap(vars(visit), nrow=2)
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

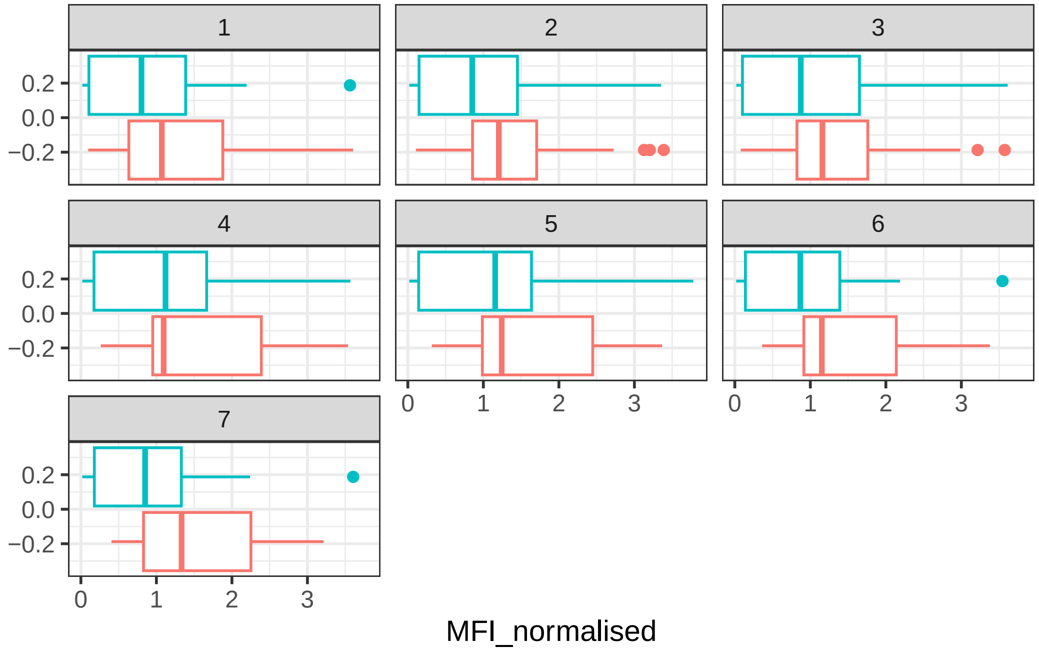


Q14. Which antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

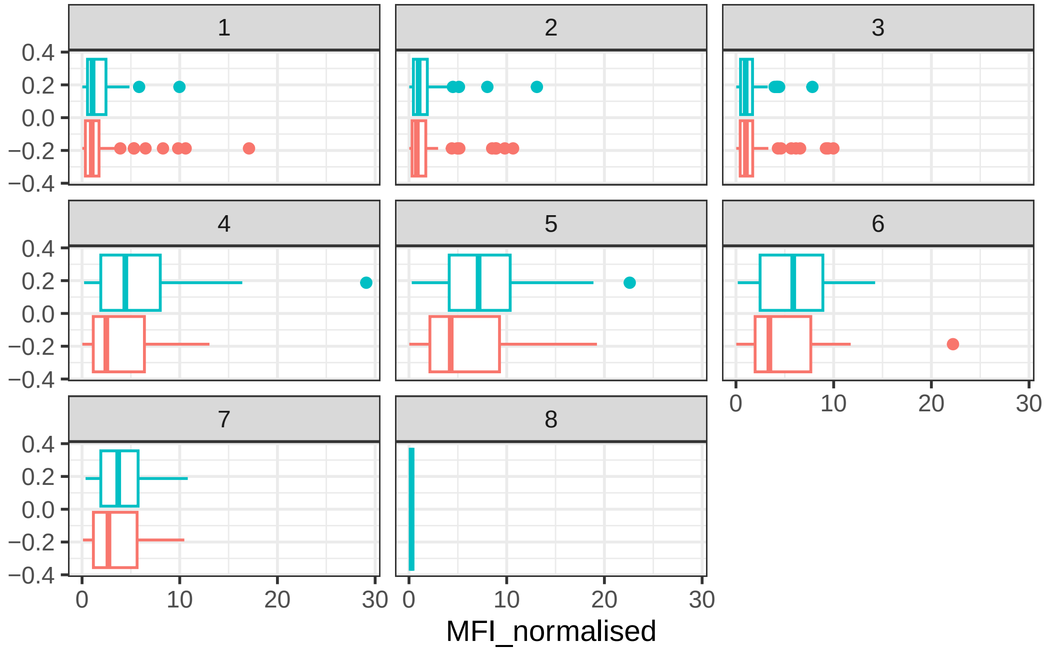
PT, FHA, PRN, and FIM2/2 - these are the antigens present in the vaccine

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).

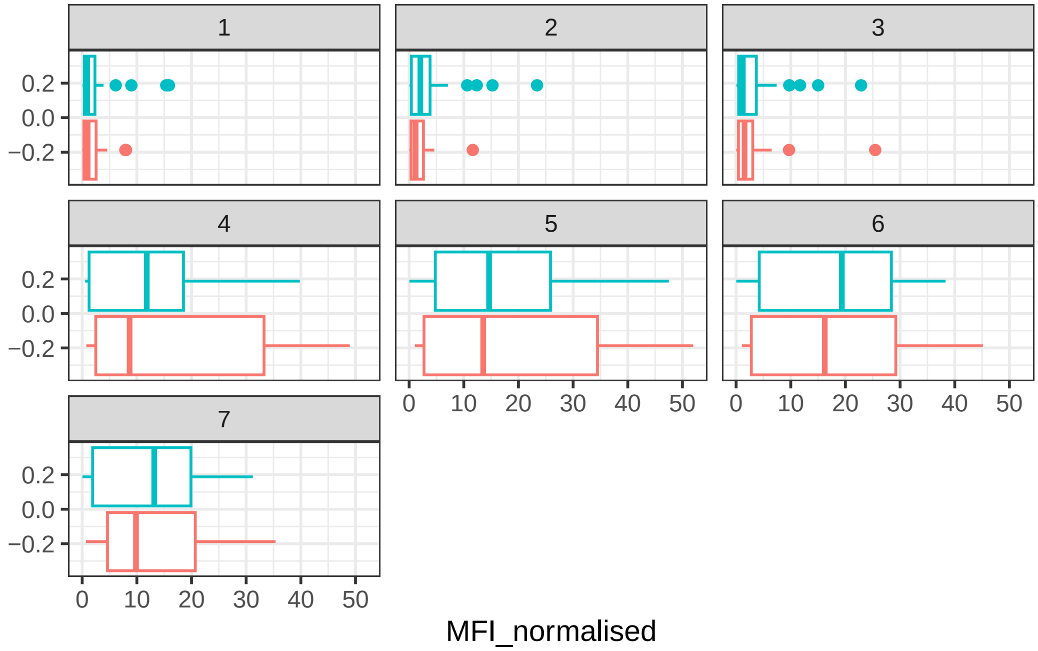
```
filter(igg, antigen=="OVA") %>%
    ggplot() +
    aes(MFI_normalised, col=infancy_vac) +
    geom_boxplot(show.legend = FALSE) +
    facet_wrap(vars(visit)) +
theme_bw()
```



```
filter(igg, antigen=="PT") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
theme_bw()
```



```
filter(igg, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
theme_bw()
```



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

The PT data shows the PT levels rise significantly comapared to OVA, which is seen with both the aP and wP vaccines.

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

Looking at the generated box plots, the differences are not apparent, whereas with the line graphs the differences are clearly visible, as the normalized MFI levels are much higher in the wP vs the aP.