

Class 19 - Mini-Project Investigating Pertussis Resurgence with CMI-PB

Hannah Kim

6/8/23

Background

Pertussis (more commonly known as whooping cough) is a highly contagious respiratory disease caused by the bacterium *Bordetella pertussis*

1. Investigating pertussis cases by year

We can view this data on the CDC website here: <https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html>.

```
# Load the necessary libraries
library(ggplot2)
#install.packages("datapasta")
library(datapasta)

# Paste your data from clipboard using datapasta
# Alternatively, you can create a data frame manually
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),
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,
6124, 2116)

```

)

Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number

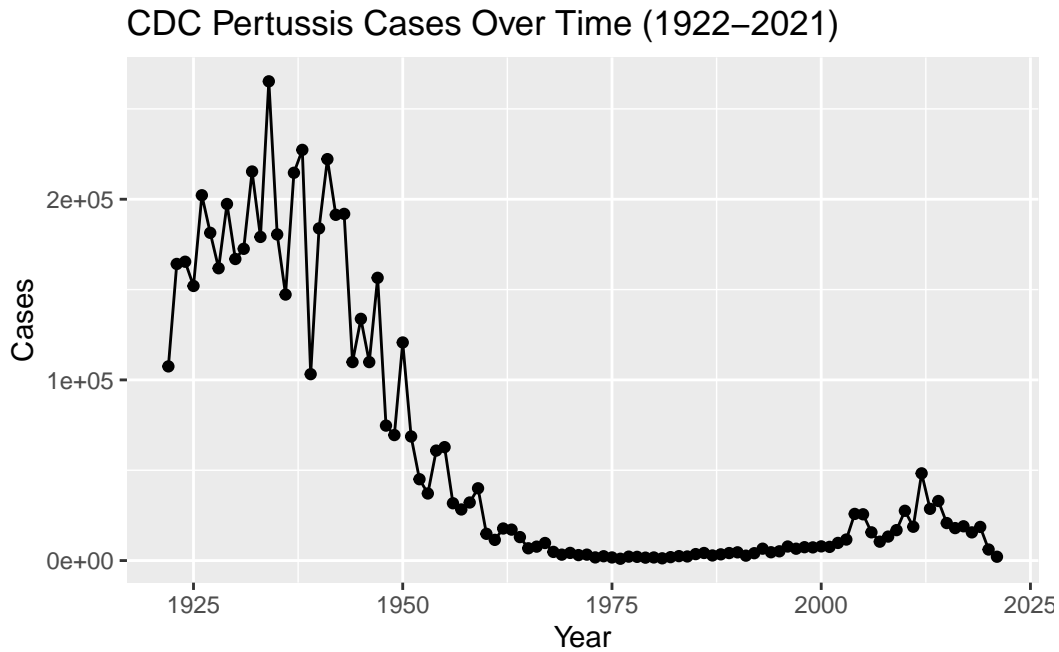
data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```

# Create the plot black and white
ggplot(cdc) + aes(x = Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +

```

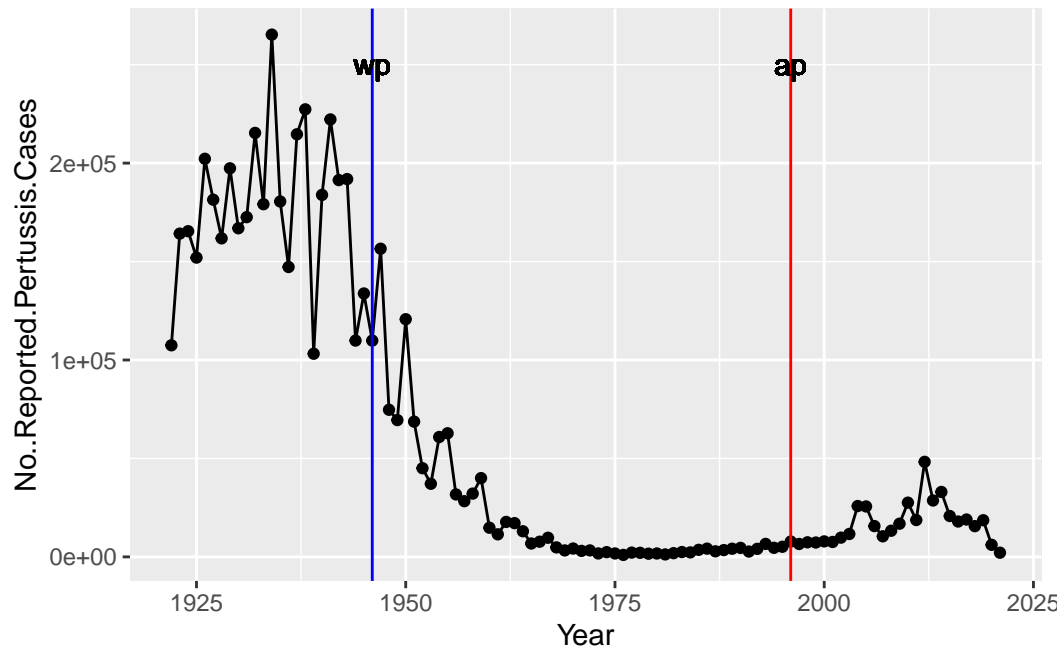
```
labs(title = "CDC Pertussis Cases Over Time (1922-2021)", x = "Year", y = "Cases")
```



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?

```
# Create the plot with labeled vaccine (wP,aP)
ggplot(cdc) + aes(x = Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 1946, col="blue") +
  geom_vline(xintercept = 1996, col="red") +
  geom_text(x=1946, y = 250000, label = "wp") +
  geom_text(x=1996, y = 250000, label = "ap")
```



```
labs(title = "CDC Pertussis Cases Over Time (1922-2021)", x = "Year", y = "Cases")
```

```
$x
```

```
[1] "Year"
```

```
$y
```

```
[1] "Cases"
```

```
$title
```

```
[1] "CDC Pertussis Cases Over Time (1922-2021)"
```

```
attr("class")
```

```
[1] "labels"
```

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

- For aP vaccine there is more sensitivity of testing for the PCR type
- People were misinformed regarding the vaccines which made them more hesitant to take them

- People were still getting vaccinated while still being developed for this data taken
- waning of immunity in adolescents originally primed as infants with the newer aP 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.

The CMI-PB API returns JSON data

```
# call upon jsonlite, allows us to work read, write Json data
library(jsonlite)
```

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/>.

```
# create a subject data frame using the read_json() function as a vector
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)
# run head to pull the top 3 data
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 infancy vaccinated subjects are in the dataset?

```
# create a table to distinguish between aP vs wP vaccine
table(subject$infancy_vac)
```

```
aP wP
47 49
```

- There is 47 aP vaccines and 49 wP vaccines

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

```
# create a table to distinguish the different sexes
table(subject$biological_sex)
```

```
Female    Male
     66     30
```

- There are 66 females and 30 males biologically

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

```
#Create a table to distinguish categories based on race and biological sex
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

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?

```
library(lubridate)
```

Attaching package: 'lubridate'

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

```
date, intersect, setdiff, union
```

```
# Assuming subject is your data frame and year_of_birth is the column with years
# calculate age
subject$age <- today() - ymd(subject$year_of_birth)
```

Next we can use dplyr's filter() function to limit ourselves to a particular subset of subjects to examine the 6 number summary of their age in years:

```
# Load the necessary libraries
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
```

```
# create a filter for the vaccine aP to pull out the infancy_vac
ap <- subject %>% filter(infancy_vac == "aP")

# use the summary function to create mean, median, and qurtiles
round( summary( time_length( ap$age, "years" ) ) )
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
23	25	26	26	26	27

```
# set a filter for vaccine wP to pull out the infancy_vac
wp <- subject %>% filter(infancy_vac == "wP")

# use the summary function to create mean, median, and qurtiles
round( summary( time_length( wp$age, "years" ) ) )
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
28	32	35	37	40	55

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

```
# Calculate the interval
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)

# Calculate age at boost in years
age_at_boost <- time_length(int, "year")

# Print the head of the age_at_boost
head(age_at_boost)
```

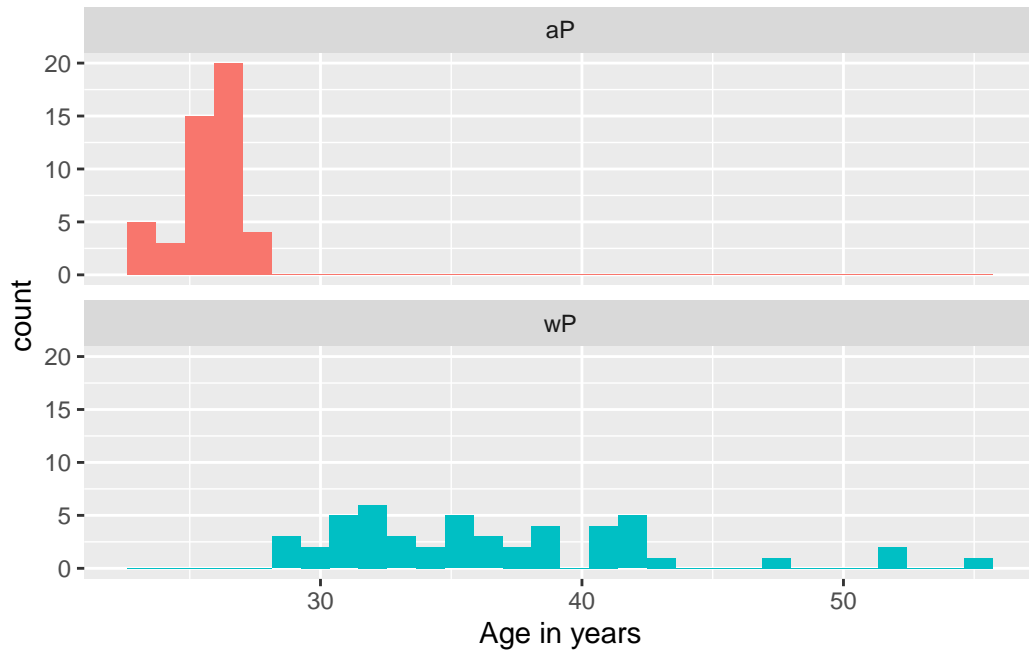
```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

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

groups are significantly different?

```
# create a ggplot using geom_histogram to form two comparable data for the aP vs wP vaccin
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`.



Statistical differences (p-value):

```
# Or use wilcox.test()
x <- t.test(time_length( wp$age, "years" ),
            time_length( ap$age, "years" ))
# print p-value
x$p.value
```

```
[1] 1.316045e-16
```

- Based on the achieved results from the ggplot and p-value for aP and wP vaccine we can determine that they are significantly different. The achieved p-value is $1.31e-16 < 0.05$

Joining multiple tables

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

```
#the API URL for specimen
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector = TRUE)
10
```

```
[1] 10
```

```
#the API URL for titer
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.

Q10. 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:

```
# use the meta as a varibale
meta <- inner_join(specimen, subject)
```

Joining with `by = join_by(subject_id)`

```
# achieve the dimmension for the meta
dim(meta)
```

```
[1] 729 14
```

```
# get the top six results for meta data
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	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	13679 days
2	13679 days
3	13679 days
4	13679 days
5	13679 days
6	13679 days

Q11. 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.

```
# Perform inner join
abdata <- inner_join(titer, meta)
```

Joining with `by = join_by(specimen_id)`

```
# Print dimensions of the resulting data frame
dim(abdata)
```

```
[1] 32675    21
```

Q12. How many specimens (i.e. entries in abdata) do we have for each isotype?

```
# create a table for the isotypes
table(abdata$isotype)
```

```

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

```

Q13. What do you notice about the number of visit 8 specimens compared to other visits?

```

# create a table for the number of visits
table(abdata$visit)

```

```

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

```

- visits 1 through 7 has more than 4000 but number 8 is only 80 specimens. each visit data is collected to run antibodies and testings
- visit 8 is significantly less than other visits by a huge margin as it could be due to ongoing sample collection for antibody testings for the year of 2021 which is when the data is capped

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.

```

# Filter the data frame
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)

# pull the top 6 data
head(ig1)

```

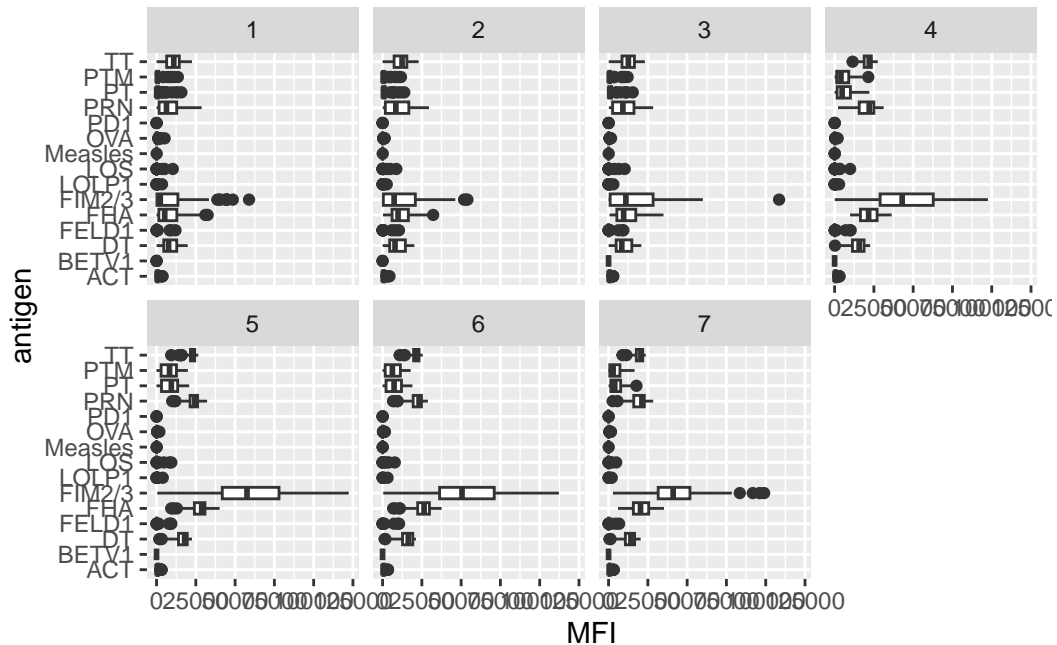
	specimen_id	isotype	is_antigen_specific	antigen	MFI	MFI_normalised
1	1	IgG1	TRUE	ACT	274.355068	0.6928058
2	1	IgG1	TRUE	LOS	10.974026	2.1645083
3	1	IgG1	TRUE	FELD1	1.448796	0.8080941
4	1	IgG1	TRUE	BETV1	0.100000	1.0000000
5	1	IgG1	TRUE	LOLP1	0.100000	1.0000000
6	1	IgG1	TRUE	Measles	36.277417	1.6638332

unit lower_limit_of_detection subject_id actual_day_relative_to_boost

1	IU/ML	3.848750	1		-3
2	IU/ML	4.357917	1		-3
3	IU/ML	2.699944	1		-3
4	IU/ML	1.734784	1		-3
5	IU/ML	2.550606	1		-3
6	IU/ML	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	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	13679 days				
2	13679 days				
3	13679 days				
4	13679 days				
5	13679 days				
6	13679 days				

Q14. Complete the following code to make a summary boxplot of Ab titer levels (MFI) for all antigens:

```
# create a ggplot using the geom_boxplot as the set parameters to display different antigens
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```

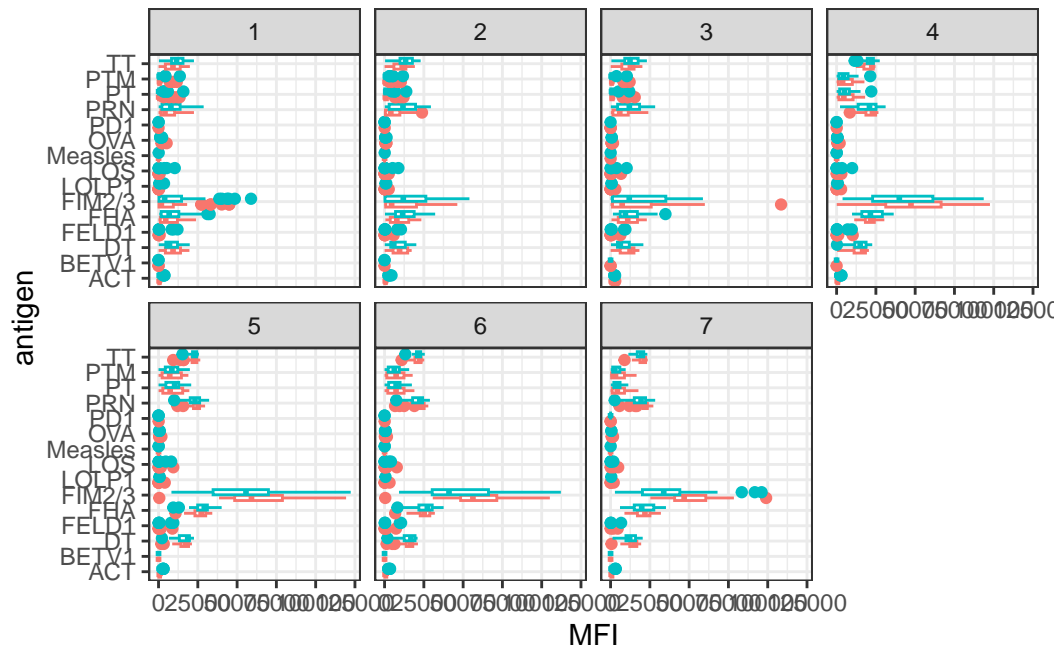


Q15. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

- Based on observing, FIM2/3 is significantly different from day 1-7 while others are constant and not changing.
- FIM2/3 is increasing each day and significantly different from the day before

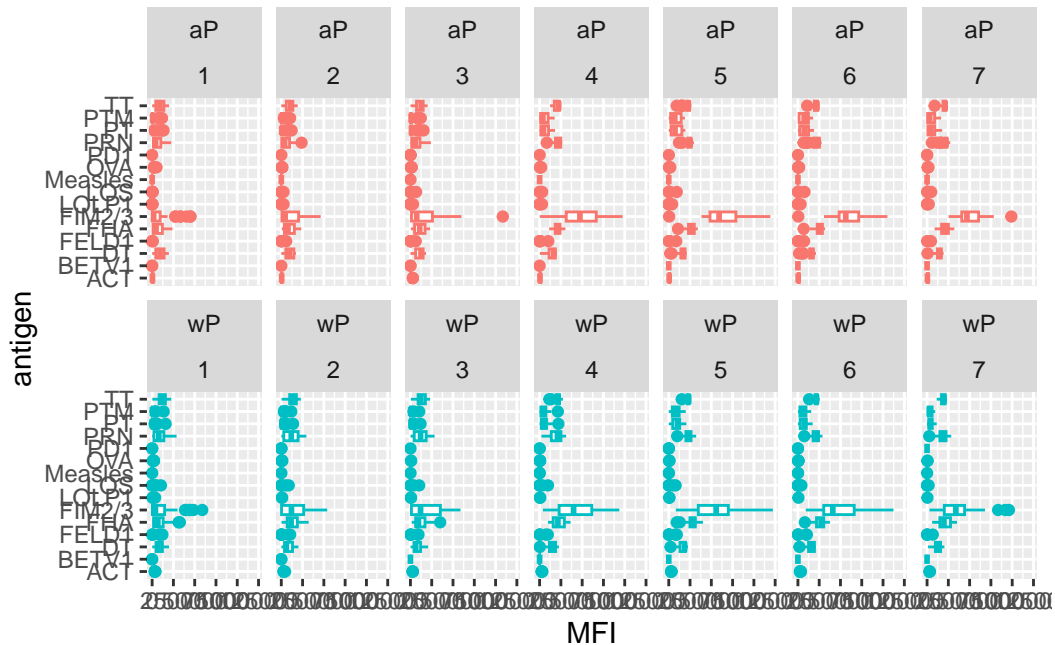
We can attempt to examine differences between wP and aP here by setting color and/or facet values of the plot to include infancy_vac status (see below)

```
# create a ggplot displaying the ig1 but instead we use color
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  theme_bw()
```



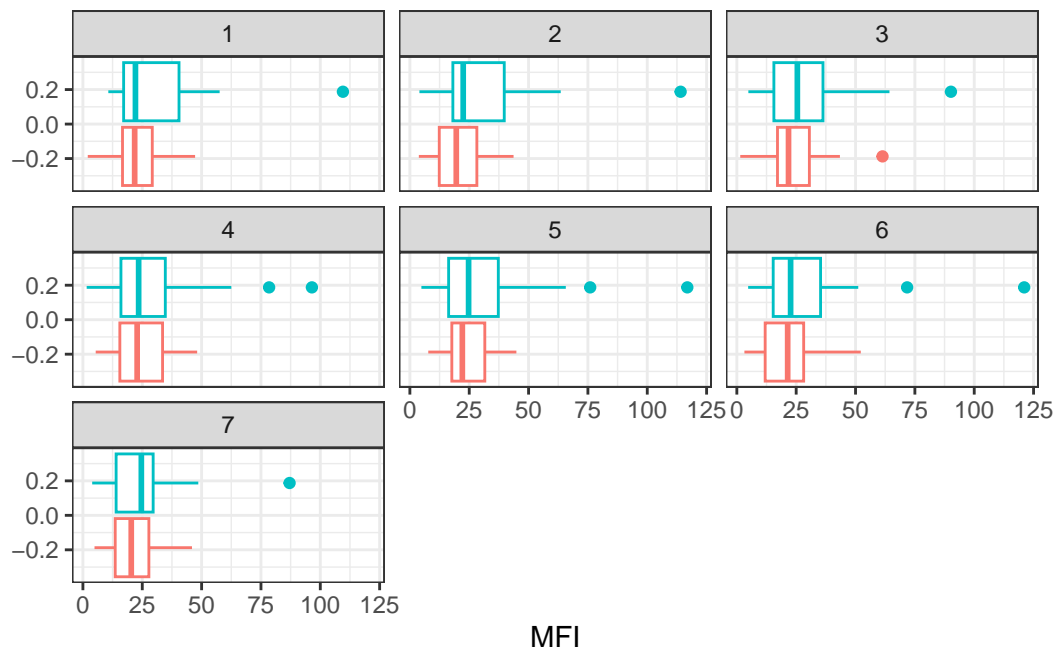
Another version of this plot adding infancy_vac to the faceting:

```
# Same ggplot but the color coordination is seperated by vaccines to create less confucion
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```



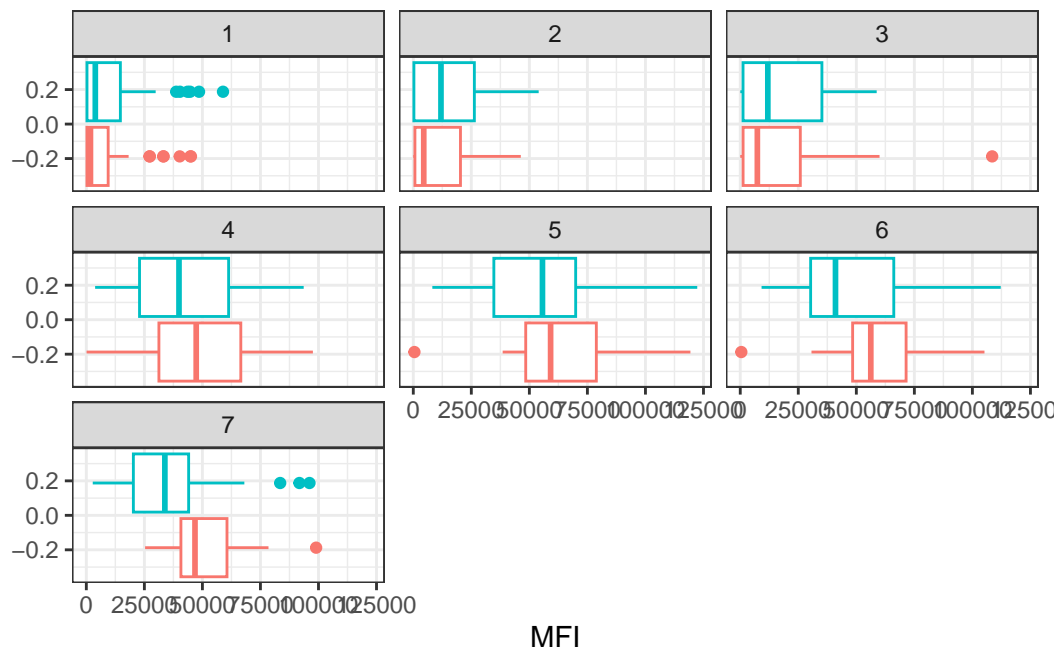
Q16. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("Measles", that is not in our vaccines) and a clear antigen of interest ("FIM2/3", extra-cellular fimbriae proteins from B. pertussis that participate in substrate attachment).

```
# Filter the data frame and create a ggplot for Measles
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```

and the same for antigen=="FIM2/3"

```
# Filter the data frame and create a ggplot for FIM2/3
filter(ig1, antigen== "FIM2/3") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



Q17. What do you notice about these two antigens time courses and the FIM2/3 data in

particular?

- Based on the produced filter boxplots FIM2/3 levels clearly rise over time and far exceed those of Measles. They also appear to peak at visit 5 and then decline. This trend appears similar for for wP and aP subjects.

Q18. Do you see any clear difference in aP vs. wP responses?

- aP and wP were both observed using the box_plot was determined to be not significantly different from each other. They seem to be increasing as each day passes by

5. Obtaining CMI-PB RNASeq data

```
# Specify the URL
#url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSNG00000211896"

# Read JSON from URL into a data frame
```

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

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)
#ssrna <- inner_join(rna, meta)
```

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

```
# create a ggplot to display the tpm per visits
#ggplot(ssrna) +
#   aes(visit, tpm, group=subject_id) +
#   geom_point() +
#   geom_line(alpha=0.2)
```

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

level)?

- The expression of the gene is at its maximum level during visit 4. it does match the trend from the previous plot where both spiked at visit 4. Visit 4 is a lot larger in both set of observation

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

- it does match the trend from the previous plot where both spiked at visit 4. Visit 4 is a lot larger in both set of observation when comparing Q15 ggplot.

We can dig deeper and color and/or facet by infancy_vac status:

```
# create a ggplot for the infancy_vac
# ggplot(ssrna) +
#   aes(tpm, col=infancy_vac) +
#   geom_boxplot() +
#   facet_wrap(vars(visit))
```

There is however no obvious wP vs. aP differences here even if we focus in on a particular visit:

```
# use the ssrna graph to create a trendline for the different vaccine types
#ssrna %>%
# filter(visit==4) %>%
# ggplot() +
# aes(tpm, col=infancy_vac) + geom_density() +
# geom_rug()
```