

Class 18: Pertussis Mini Project

Anyoleth Alarcon A17347293

Table of contents

Background	1
Examining cases of Pertussis by year	1
Enter the CMI-PB project	3
Focus on IgG levels	11

Background

Pertussis (aka Whooping Cough) is a common lung infection caused by the bacteria *B. Pertussis*.

The CDC tracks cases of Pertussis in the US: <https://www.cdc.gov/pertussis/php/surveillance/pertussis-cases-by-year.html>

Examining cases of Pertussis by year

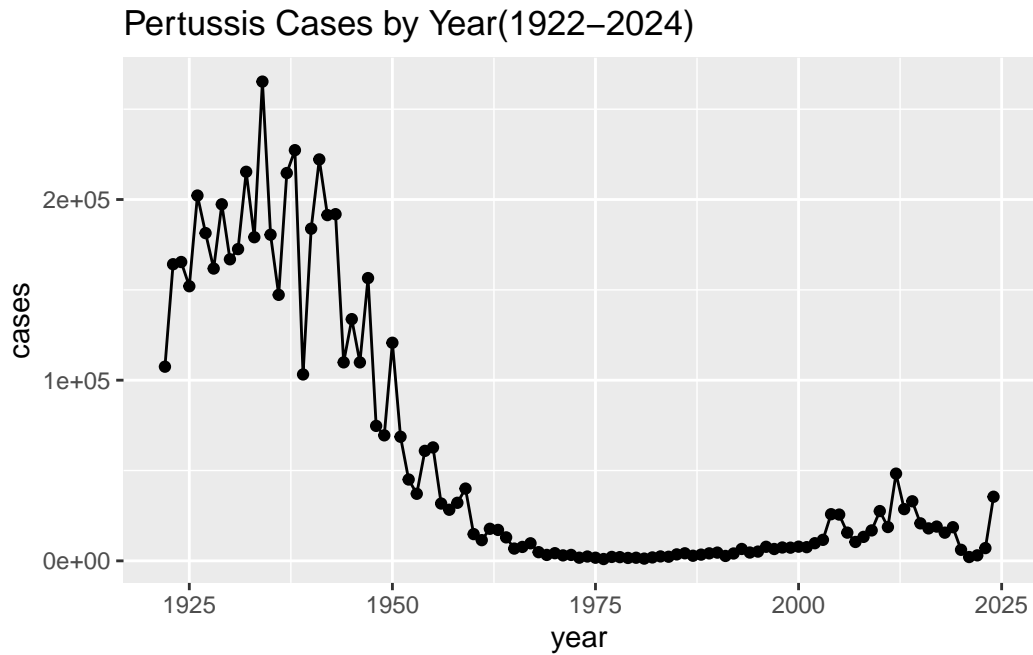
We can use the **datapasta** package to scrape case numbers of

Q. Make a plot of pertussis cases per year using ggplot.

```
library(ggplot2)

cases <- ggplot(cdc, aes(year, cases))+
  geom_point() +
  geom_line() +
  labs(title = "Pertussis Cases by Year(1922-2024)")

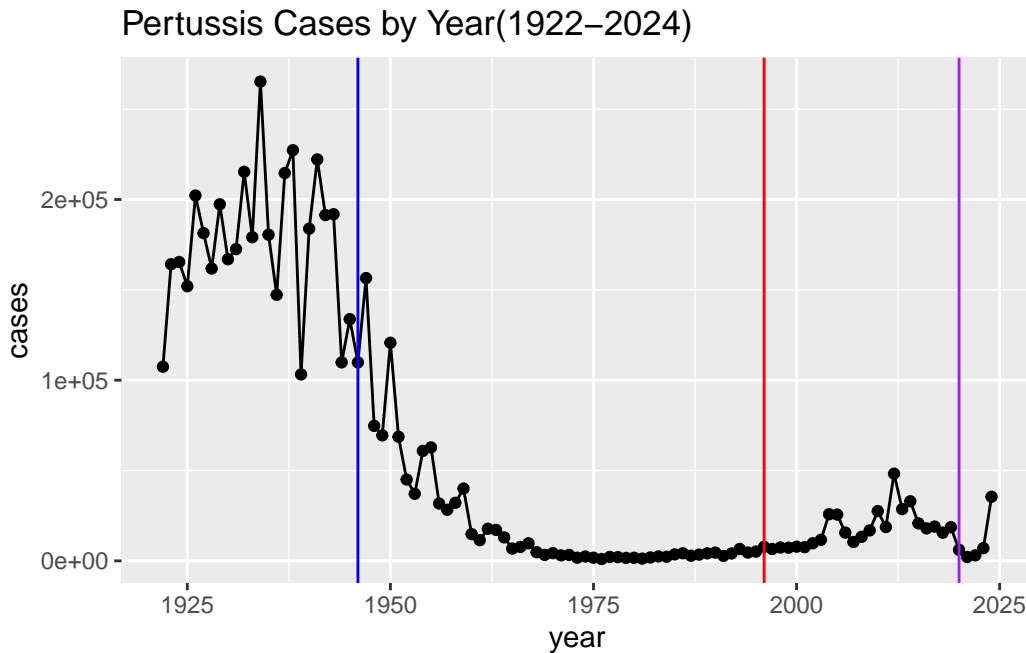
cases
```



Q2. Add some key time points in our history of interaction with Pertussis. These include wP roll-out (the first vaccine) in 1946 and the switch to aP in 1996.

We can use `geom_vline()` for this.

```
cases + geom_vline(xintercept = 1946, col="blue") +  
  geom_vline(xintercept = 1996, col="red") +  
  geom_vline(xintercept = 2020, col="purple")
```



After the first roll-out vaccine (blue line) the cases for Whooping Cough decreases in the following years. After the switch to aP (red line) in 1996, do see a rise in cases, and can be due to different factors. Such as vaccine hesitancy, or bacteria vaccine immunity, or effectivity. Mounting evidence suggests that the newer **aP** vaccine is less effective over the long term than the older **wP** vaccine that it replaced. IN other words, vaccine protection wanes more rapidly with aP than with wP.

Enter the CMI-PB project

CMI-PB (Computational Models of Immunity - Pertussis Boost) major goal is to investigate how the immune system responds differently to with aP vs wP vaccinated individuals and be able to predict this at an early age.

CMB_PB makes all their collected data freely available and they store it in a database composed different tables. Here we will access a few of these.

We can use the **jsonlite** package to read this data

```
library(jsonlite)

subject <- read_json("https://www.cmi-pb.org/api/v5_1/subject", simplifyVector = TRUE)

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

Q. How many subjects (i.e. enrolled people) are there in this dataset?

```
nrow(subject)
```

```
[1] 172
```

Q. How many “aP” and “wP” subjects are there?

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

```
aP wP
87 85
```

Q. How many Male/Female are in the dataset?

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

```
Female  Male
112     60
```

Q. How about gender and race numbers?

```
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

Q. Is this representative of the US population?

Naur.

Let's read another database table from CMI-PB

```
specimen <- read_json("https://www.cmi-pb.org/api/v5_1/specimen", simplifyVector = TRUE)
```

```
ab_data <- read_json("https://www.cmi-pb.org/api/v5_1/plasma_ab_titer", simplifyVector = TRUE)
```

wee peak at these

```
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

```
head(ab_data)
```

	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

We want to “join” these tables to get all our information together. For this we will use **dplyr** package and the `inner_join()` function.

```
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(subject, specimen)
```

Joining with `by = join_by(subject_id)`

```
head(meta)
```

	subject_id	infancy_vac	biological_sex	ethnicity	race
1	1	wP	Female Not Hispanic or Latino	White	
2	1	wP	Female Not Hispanic or Latino	White	
3	1	wP	Female Not Hispanic or Latino	White	
4	1	wP	Female Not Hispanic or Latino	White	
5	1	wP	Female Not Hispanic or Latino	White	
6	1	wP	Female Not Hispanic or Latino	White	

	year_of_birth	date_of_boost	dataset	specimen_id
1	1986-01-01	2016-09-12	2020_dataset	1
2	1986-01-01	2016-09-12	2020_dataset	2
3	1986-01-01	2016-09-12	2020_dataset	3
4	1986-01-01	2016-09-12	2020_dataset	4
5	1986-01-01	2016-09-12	2020_dataset	5
6	1986-01-01	2016-09-12	2020_dataset	6

	actual_day_relative_to_boost	planned_day_relative_to_boost	specimen_type
1	-3	0	Blood
2	1	1	Blood
3	3	3	Blood
4	7	7	Blood
5	11	14	Blood
6	32	30	Blood

	visit
1	1
2	2
3	3
4	4
5	5
6	6

One more “join” to get ab_data and meta all together

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

Joining with `by = join_by(specimen_id)`

```
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 subject_id infancy_vac biological_sex						
1	UG/ML	2.096133	1	wP	Female	
2	IU/ML	29.170000	1	wP	Female	
3	IU/ML	0.530000	1	wP	Female	
4	IU/ML	6.205949	1	wP	Female	
5	IU/ML	4.679535	1	wP	Female	
6	IU/ML	2.816431	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	
actual_day_relative_to_boost planned_day_relative_to_boost specimen_type						
1		-3		0	Blood	
2		-3		0	Blood	
3		-3		0	Blood	
4		-3		0	Blood	
5		-3		0	Blood	
6		-3		0	Blood	
visit						
1	1					
2	1					
3	1					
4	1					
5	1					
6	1					

```
dim(abdata)
```

```
[1] 61956    20
```

Q. How many Ab isotypes are there in the dataset?


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

```
  IgE   IgG  IgG1  IgG2  IgG3  IgG4  
6698  7265 11993 12000 12000 12000
```

Q. How many different antigens are measured in the dataset?

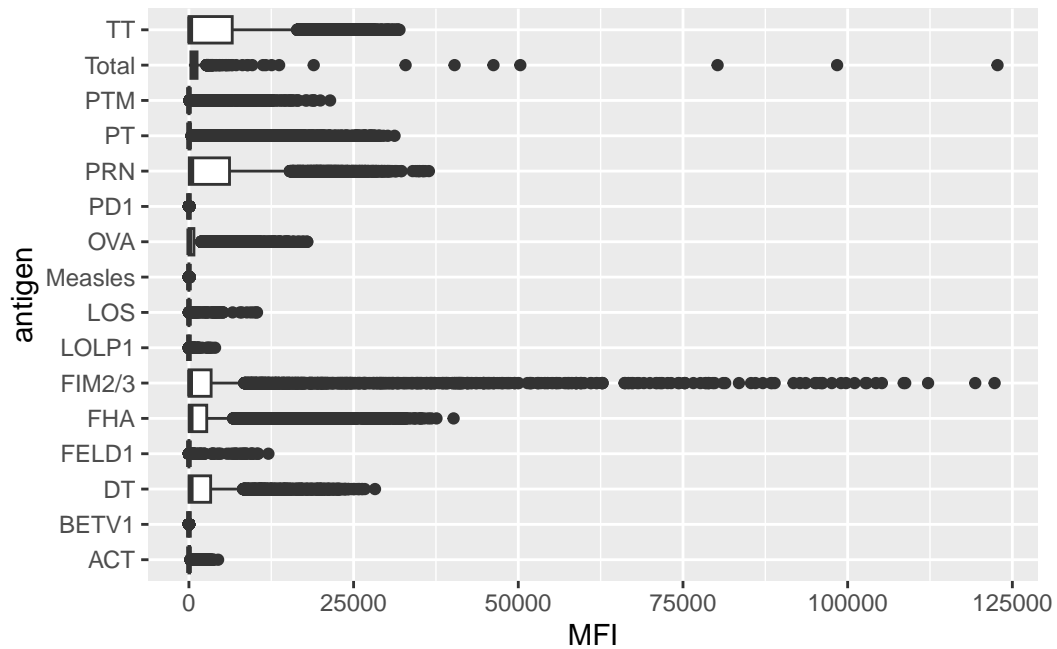
```
table(abdata$antigen)
```

```
  ACT  BETV1    DT  FELD1    FHA  FIM2/3  LOLP1    LOS Measles    OVA  
1970   1970   6318   1970   6712   6318   1970   1970   1970   6318  
  PD1    PRN    PT    PTM  Total    TT  
1970   6712   6712   1970    788   6318
```

Q. Make a boxplot of antigen levels across the whole dataset (MFI vs antigen).

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

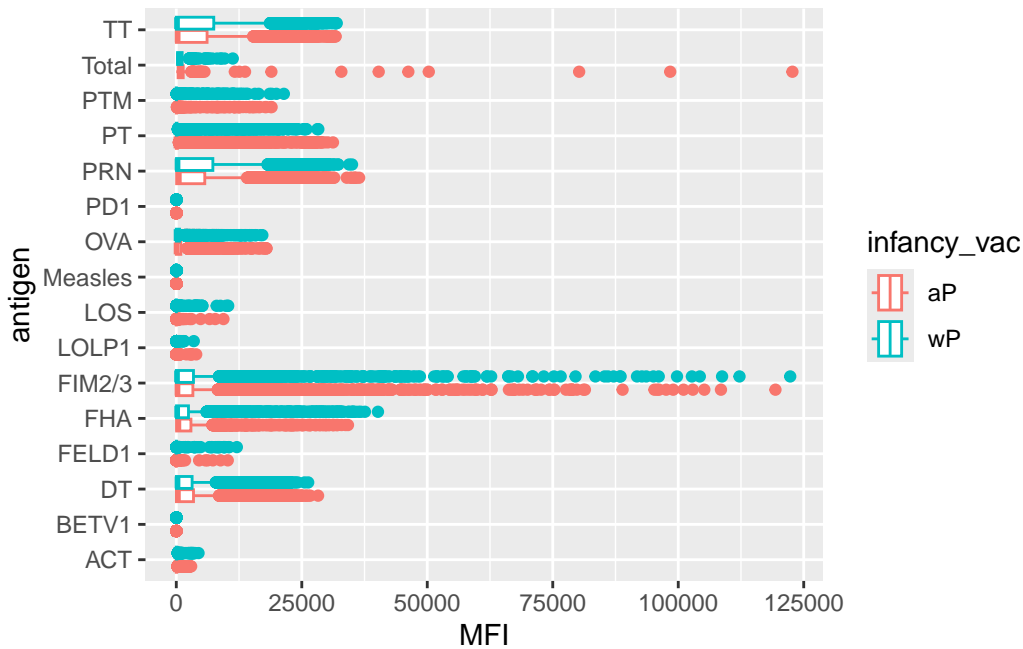
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).



Q. Are there obvious differences between aP and wP values?

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

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



Focus on IgG levels

IgG is the most abundant antibody in blood. With four sub-classes (IgG1 to IgG4) crucial for long-term immunity and responding to bacterial & viral infections.

```
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	infancy_vac	biological_sex	
1	IU/ML	0.530000	1	wP	Female	
2	IU/ML	6.205949	1	wP	Female	
3	IU/ML	4.679535	1	wP	Female	
4	IU/ML	0.530000	3	wP	Female	
5	IU/ML	6.205949	3	wP	Female	
6	IU/ML	4.679535	3	wP	Female	

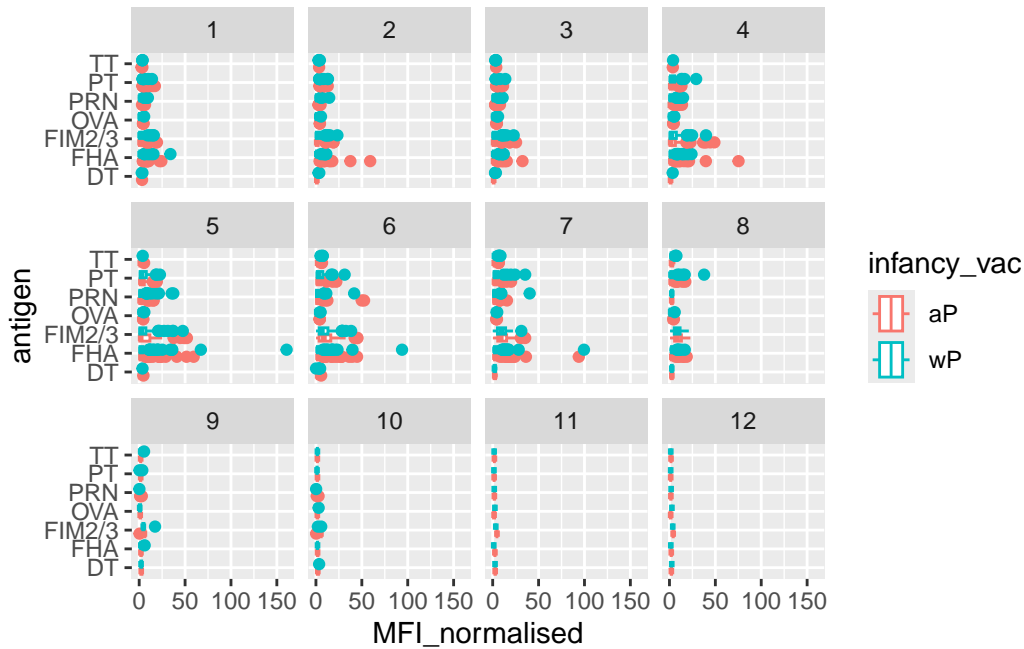
	ethnicity	race	year_of_birth	date_of_boost	dataset
1	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
2	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
3	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
4	Unknown	White	1983-01-01	2016-10-10	2020_dataset
5	Unknown	White	1983-01-01	2016-10-10	2020_dataset
6	Unknown	White	1983-01-01	2016-10-10	2020_dataset

	actual_day_relative_to_boost	planned_day_relative_to_boost	specimen_type
1	-3	0	Blood
2	-3	0	Blood
3	-3	0	Blood
4	-3	0	Blood
5	-3	0	Blood
6	-3	0	Blood

	visit
1	1
2	1
3	1
4	1
5	1
6	1

Same boxplot of antigens as before

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



Focus in further in just one of these antigens - let's pick **PT** (Pertussis Toxin, one of the main toxins of the bacteria) in the **2021_dataset** again for **IgG** antibody isotypes.

```
table(igg$dataset)
```

```
2020_dataset 2021_dataset 2022_dataset 2023_dataset
      1182       1617       1456       3010
```

```
pt_igg <- abdata |>
  filter(isotype == "IgG",
         antigen == "PT",
         dataset == "2021_dataset")
```

```
dim(pt_igg)
```

```
[1] 231  20
```

```
ggplot(pt_igg,
       aes(actual_day_relative_to_boost, MFI_normalised,
          col=infancy_vac,
```

```

      group=subject_id)
    ) +
    geom_point() +
    geom_line() +
    theme_bw() +
    geom_vline(xintercept = 0) +
    geom_vline(xintercept = 14)

```

