Pertussis Mini-Project

Andy Hsu

Pertussis is a bacterial infection that causes closing of the airways and a severe cough. This mini-project will examine some of the data surrounding this disease which has recently made a resurgence.

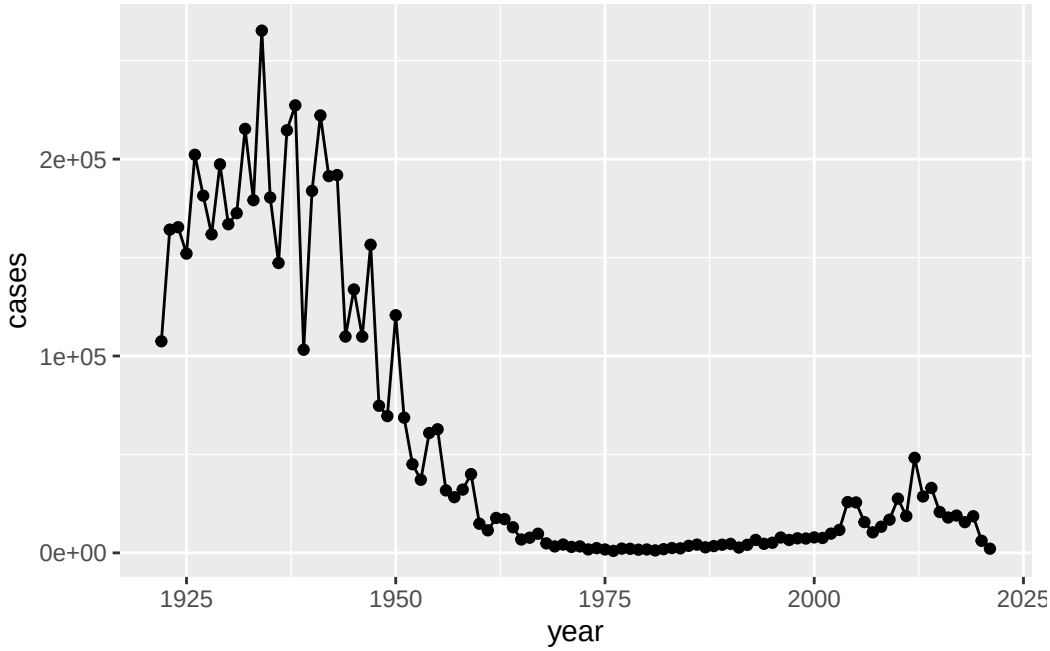
Cases by Year

We can visit this link to find data on yearly infection rates.

Now, let's plot the cases by years via **ggplot2**.

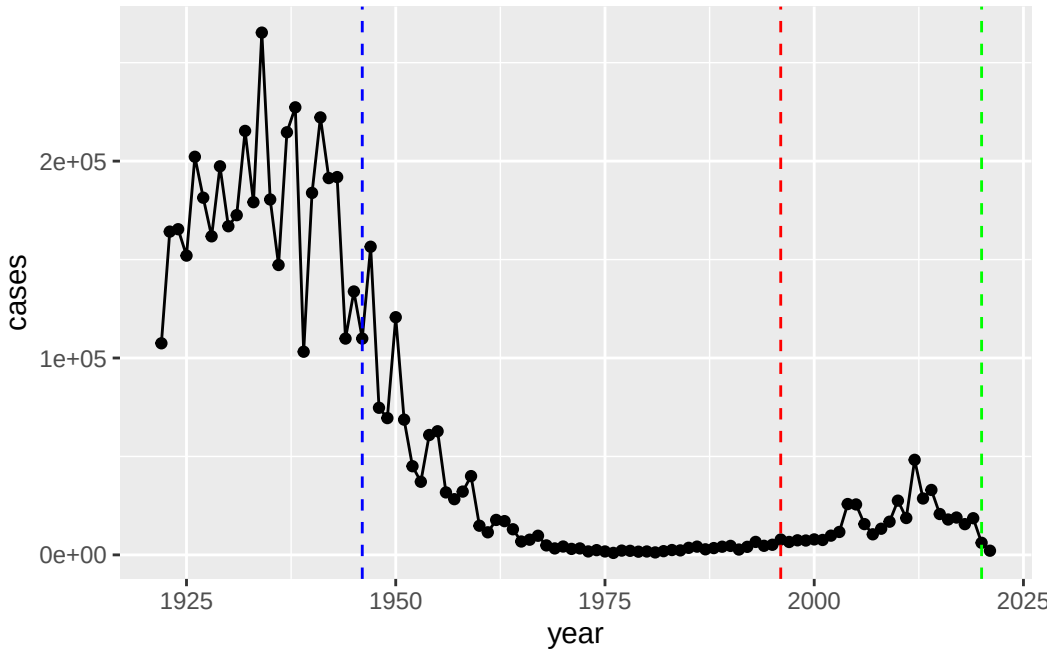
```
library(ggplot2)

cases <- ggplot(cdc, aes(year,cases)) +
  geom_point(col="black") +
  geom_line(col="black")
cases
```



Major milestones in the pertussis vaccine timeline are the introduction of the wP vaccine in 1946 and the switch to the aP vaccine in 1996. Let's add these points to the plot to view their effects. While we're at it, we can also include the Covid-19 pandemic in 2020.

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



It appears from this graph that after the switch to the aP vaccine, a large rise in cases was seen, possibly due to the aP vaccine being much less effective than the wP.

CMI-PB Data

The CMI-PB project aims to solve this problem by studying the long-term immune effects of individuals taken wP or aP. This data is documented and available on their site here.

Notice that the data stored on this site is in the JSON file format. To read this data, we will use the package **jsonlite**.

```
library(jsonlite)
```

```
Warning: package 'jsonlite' was built under R version 4.3.2
```

```
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector=T)
specimen <- read_json("http://cmi-pb.org/api/specimen", simplifyVector=T)
titer <- read_json("http://cmi-pb.org/api/v4/plasma_ab_titer", simplifyVector=T)
```

Taking a quick look at the data, we can see some distributions of subject demographics. Interestingly, there is a disproportionally large sample of Asians, Caucasians, and females, which is perhaps not very representative of the overall US demographic.

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

```
aP wP
60 58
```

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

```
Female   Male
    79     39
```

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

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

Another aspect of this data that we can examine is the age of subjects, having a correlation with immune response. Using the **lubridate** package, we can work with days extremely easily.

```
library(lubridate)
```

```
Warning: package 'lubridate' was built under R version 4.3.2
```

```
today()
```

```
[1] "2023-12-07"
```

```
  mdy("11-28-2001")
```

```
[1] "2001-11-28"
```

```
  today() - mdy("11-28-2001")
```

```
Time difference of 8044 days
```

```
  time_length( today() - mdy("11-28-2001"), "years" )
```

```
[1] 22.02327
```

Using these functions, we can calculate the average ages for wP and aP individuals, and see that the difference is

```
  subject$age <- time_length( today() - ymd( subject$year_of_birth ), "years" )
  mean( subject$age[subject$infancy_vac=="aP"] )
```

```
[1] 26.0303
```

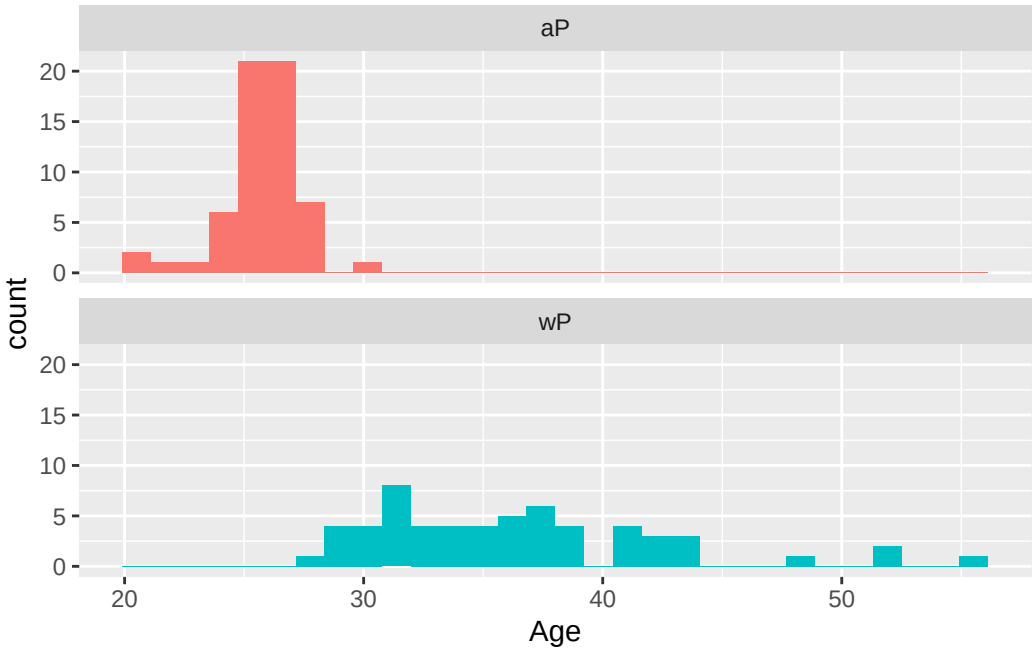
```
  mean( subject$age[subject$infancy_vac=="wP"] )
```

```
[1] 36.32703
```

Now, let's plot a histogram of the age distribution.

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

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



These graphs still need to be linked together by `subject_id`, which we can do with the `full_join()` function from 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
```

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

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

Finally, let's join titer data to the previously joined data frame.

```
datafull <- inner_join(joined,titer)
```

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

By tabling the isotypes documented in this full dataset, we can see the distributions of antibodies observed by this study.

```
table(datafull$isotype)
```

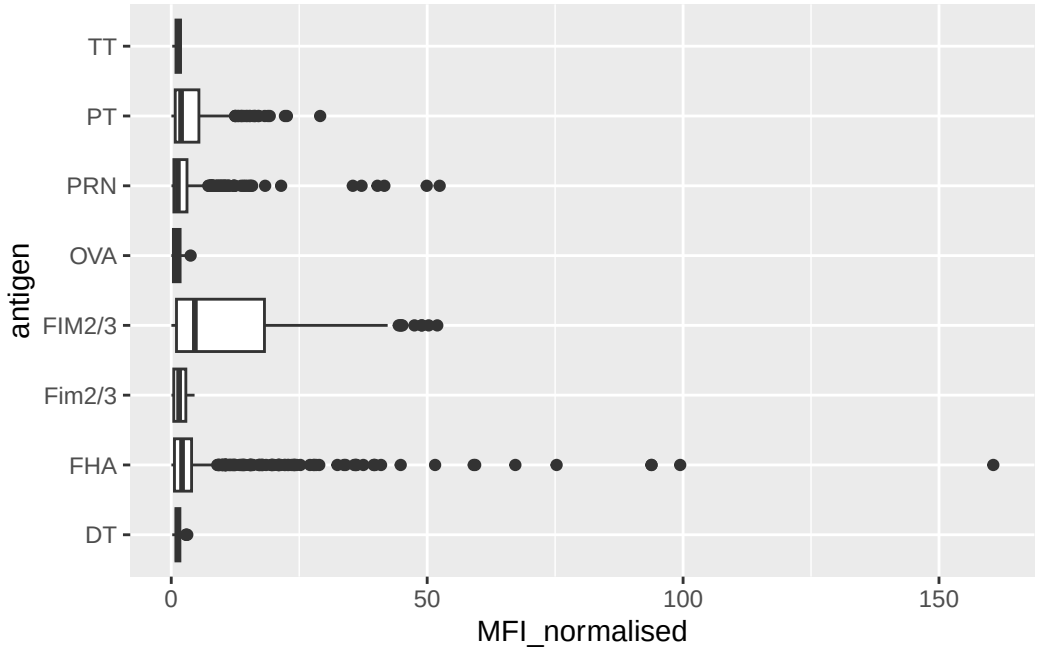
```
 IgE  IgG IgG1 IgG2 IgG3 IgG4
6698 3240 7968 7968 7968 7968
```

We can focus on one of these, IgG, and filter a new data frame.

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

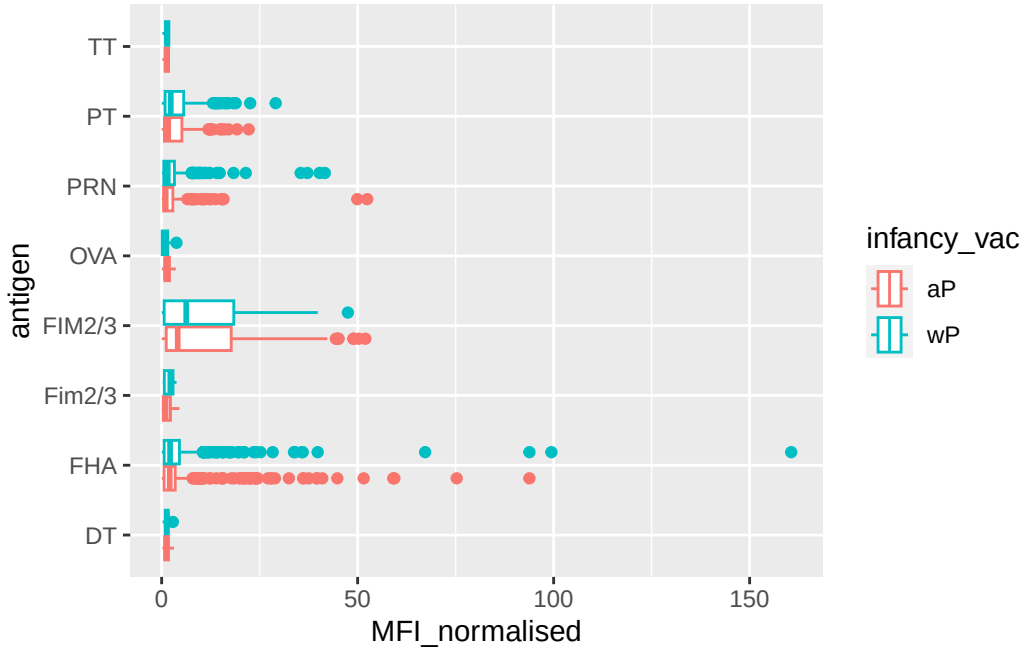
Next, we can graph a box plot of these IgG values by MFI_normalised.

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



We can also separate the two vaccine types to compare them.

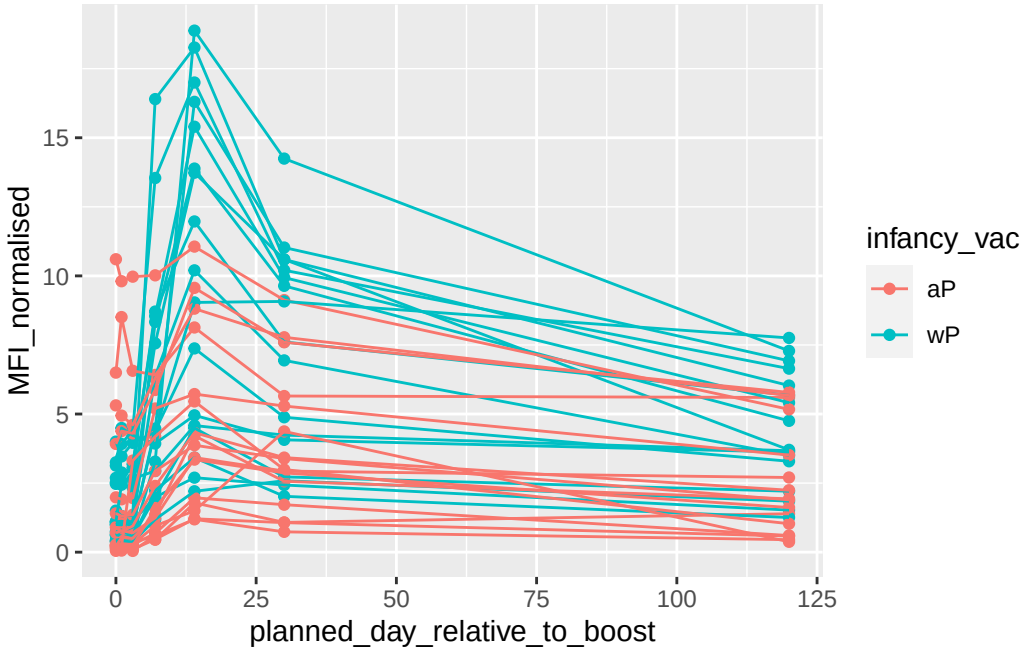
```
ggplot(igg, aes(MFI_normalised,antigen,col=infancy_vac)) +
  geom_boxplot()
```



Let's plot one last graph focusing on IgG to pertussis toxin (PT) antigen in the 2021 dataset.

```
igg.pt <- igg %>% filter(antigen=="PT", dataset=="2021_dataset")

ggplot(igg.pt, aes(planned_day_relative_to_boost, MFI_normalised, col=infancy_vac, group=s
  geom_point() +
  geom_line()
```



And that's it for this exploration of pertussis cases and the CMI-PB database.