

Pertussis and the CMI-PB project

Kendrick Nguyen

Pertussis is a bacterial infection that causes a severe cough. Often named “whooping cough”

Look at case numbers of pertussis in the US. We can use this website (“<https://www.cdc.gov/pertussis/surveillance/cases-by-year.html>”)

```
library(datapasta)
# echo: FALSE
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.

```

library(ggplot2)
library(tidyverse)

```

```

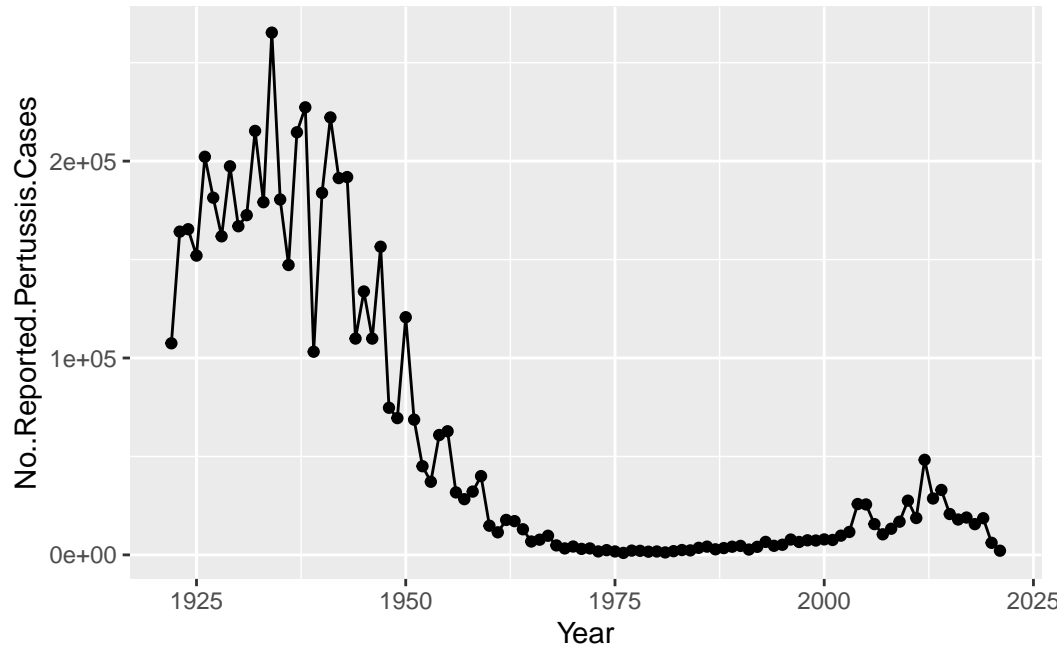
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.1.3      v readr      2.1.4
v forcats    1.0.0      v stringr    1.5.1
v lubridate  1.9.3      v tibble     3.2.1
v purrr      1.0.2      v tidyr      1.3.0
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become

```

```

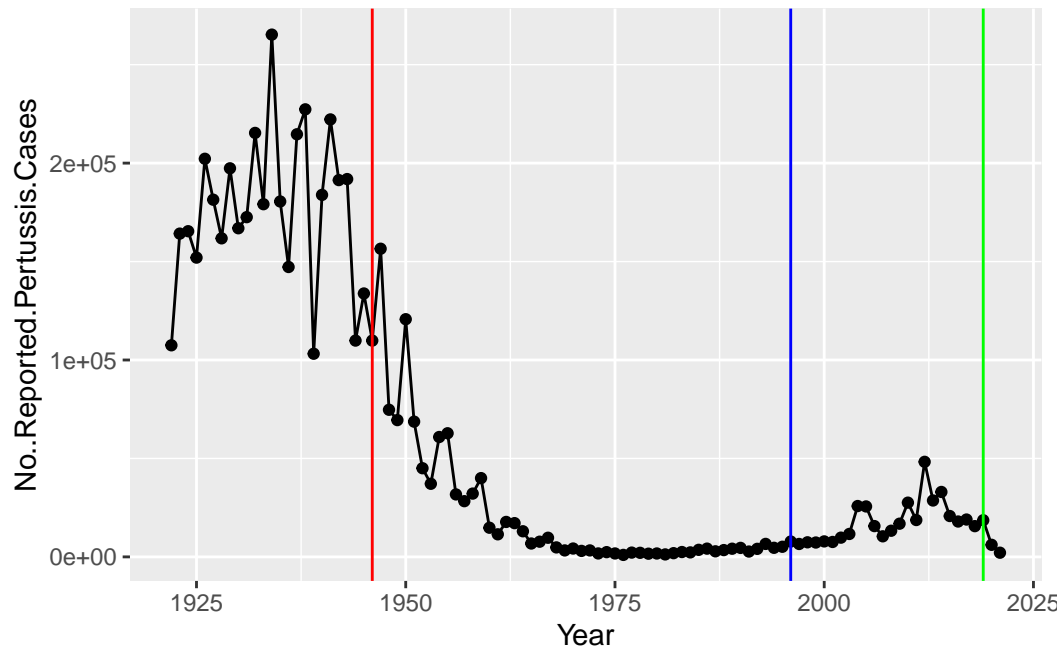
ggplot(cdc) +
  aes(Year,No..Reported.Pertussis.Cases ) +
  geom_point() +
  geom_line()

```



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?

```
ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases ) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 1946, color="red") +
  geom_vline(xintercept = 1996, color= "blue") +
  geom_vline(xintercept = 2019, color= "green")
```



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

cases of whooping cough started to rise meaning that people are losing immunity to the disease. This could mean that the new aP vaccine is not as robust as the last vaccine or the bacteria is gaining resistance to the vaccines. There could be a lot of people not getting vaccinated too.

#CMI-PB project

The CMI-PB project collects and makes available data on the immune response to pertussis booster vaccination

```
library(jsonlite)
```

Attaching package: 'jsonlite'

The following object is masked from 'package:purrr':

```
flatten
```

```
subject <- read_json("https://www.cmi-pb.org/api/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

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

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

```
aP wP
60 58
```

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

```
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

```
today() - mdy("08-01-1997")
```

Warning in with_tz.default(Sys.time(), tzzone): Unrecognized time zone ''

Time difference of 9624 days

```
time_length( today() - mdy("08-01-1997"), "years")
```

Warning in with_tz.default(Sys.time(), tzzone): Unrecognized time zone ''

```
[1] 26.34908
```

```
library(lubridate)
```

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?

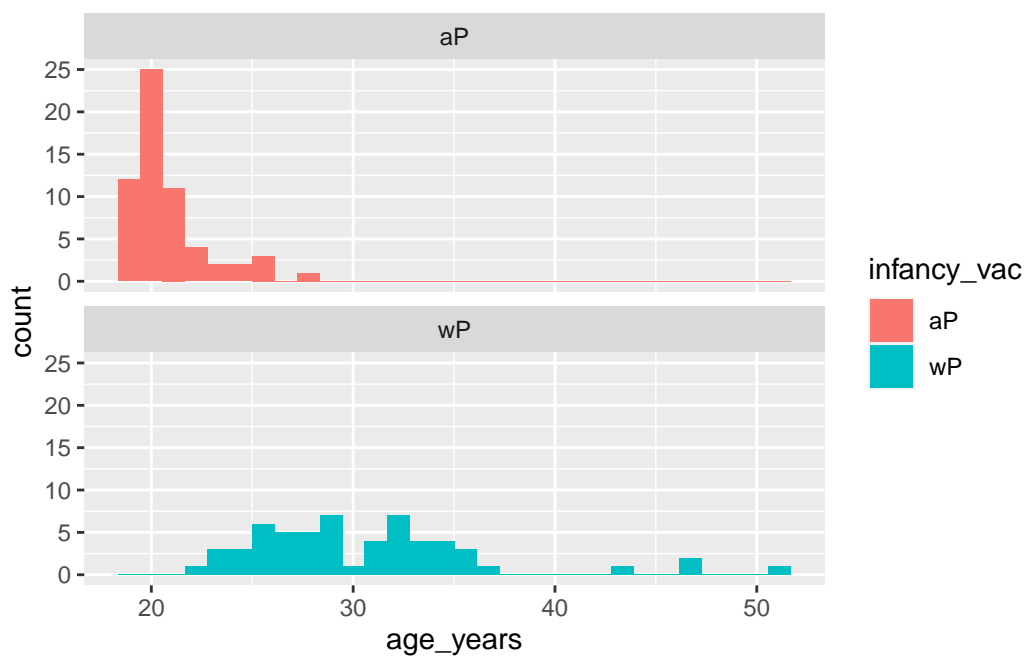
```
subject$age <-ymd(subject$date_of_boost)-ymd(subject$year_of_birth)
```

```
subject$age_years<- (time_length(subject$age, "years"))
```

```
ggplot(subject) +
  aes(age_years, fill= infancy_vac)+
  facet_wrap(vars(infancy_vac), ncol=1)+
```

```
geom_histogram()
```

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



```
table(subject$dataset)
```

```
2020_dataset 2021_dataset 2022_dataset
        60             36             22
```

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

I want to merge the specimen and subject tables together

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

Joining with `by = join_by(subject_id)`

```
dim(meta)
```

```
[1] 939 15
```

```
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 age_years
1 11212 days 30.69678
2 11212 days 30.69678
3 11212 days 30.69678
4 11212 days 30.69678
5 11212 days 30.69678
6 11212 days 30.69678

```

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.

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

Joining with `by = join_by(specimen_id)`

```
dim(abdata)
```

```
[1] 41810    22
```

```
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 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

	age	age_years
1	11212 days	30.69678
2	11212 days	30.69678
3	11212 days	30.69678
4	11212 days	30.69678
5	11212 days	30.69678
6	11212 days	30.69678

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

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

```

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

```

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

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

```
2020_dataset 2021_dataset 2022_dataset
      31520      8085      2205
```

```
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

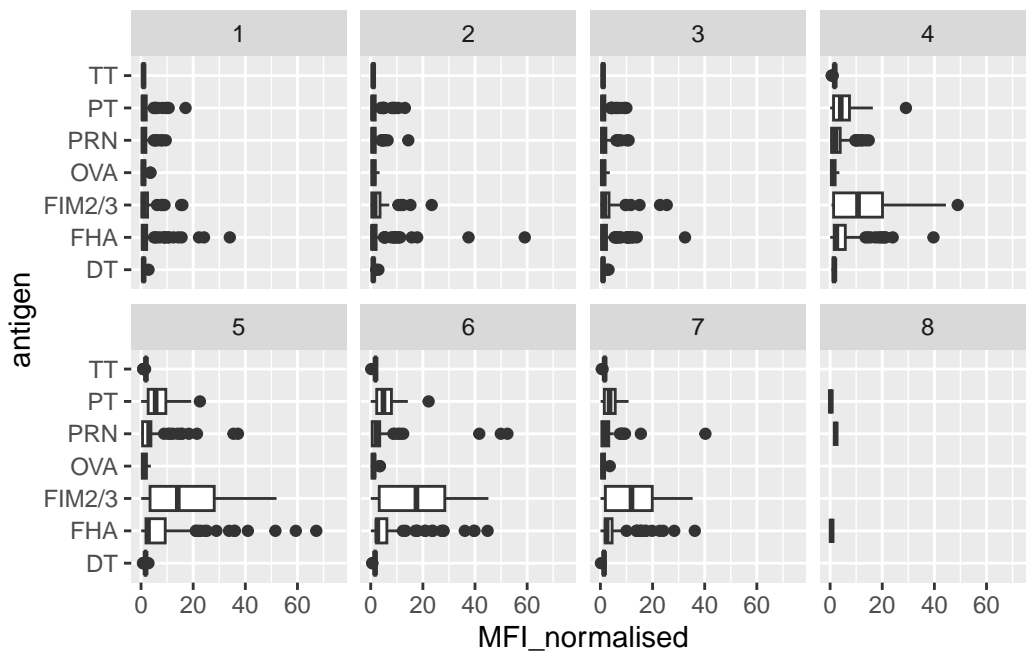
	ethnicity	race	year_of_birth	date_of_boost	dataset
1	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
2	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
3	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
4	Unknown	White	1983-01-01	2016-10-10	2020_dataset
5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

	age	age_years
1	11212 days	30.69678
2	11212 days	30.69678
3	11212 days	30.69678
4	12336 days	33.77413
5	12336 days	33.77413
6	12336 days	33.77413

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

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

Warning: Removed 5 rows containing non-finite values (`stat_boxplot()`).



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

PT, PRN, FIM 2/3, FHA

```
opps<- abdata %>% filter (antigen=="Fim2/3")
table(opps$dataset)
```

< table of extent 0 >

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

2020_dataset	2021_dataset	2022_dataset
31520	8085	2205

select for the 2021 dataset and isotype igg I want a time course of (planned_day_relative_to_boost)
IgG MFI_Normalised) for “pT” antigen

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

ggplot(results) +
  aes(planned_day_relative_to_boost, MFI_normalised, col=infancy_vac) +
  geom_point()+
  geom_line(aes(group=subject_id), linewidth= 0.5, alpha=0.5)+
  geom_smooth(se=FALSE, span =0.4, linewidth=3)
```

`geom_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -0.6

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 3.6

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 1.8382e-16

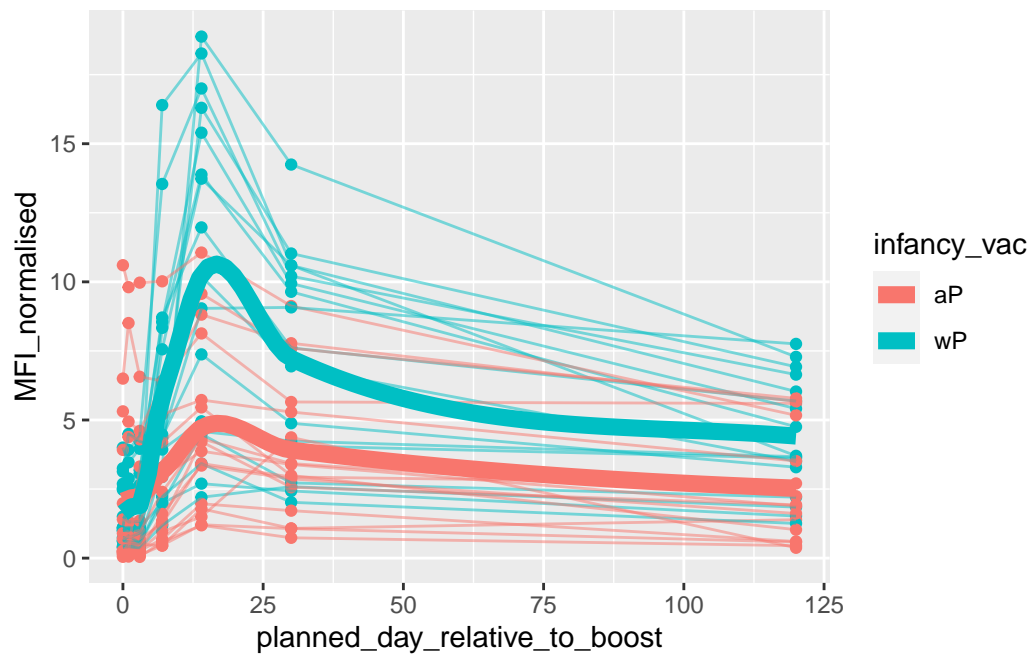
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 11364

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -0.6

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 3.6

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 1.4316e-16

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 11364
```



```
results1<- filter(abdata, dataset=="2022_dataset",antigen=="PT", isotype=="IgG")

ggplot(results1) +
  aes(planned_day_relative_to_boost, MFI_normalised, col=infancy_vac) +
  geom_point()+
  geom_line(aes(group=subject_id), linewidth= 0.5, alpha=0.5)+
  geom_smooth(se=FALSE, span =0.4, linewidth=3)
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -30.15
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 15.15
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0
```

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 229.52

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -30.15

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 15.15

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 229.52

