Class 18: Mini Project - Investigating Pertussis Resurgence

Cynthia Perez (A16393492)

Pertussis (whopping cough) is a highly contagious lung infection that is most severe in the very young (under 1 year of age)

Investigating Pertussis Cases by Year

We start with looking at Pertussis case numbers per year in the United States from the CDC website.

> 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.

First install datapasta package in console using `install.packages()` Go to https://www.cdc.gov/pertussis/surv reporting/cases-by-year.html to copy table data onto clipboard Then use Paste as data.frame to transfer the data from the clipboard into a dataframe.

```
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),
  Reported.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)
)
```

Using ggplot to make a plot of cases over time

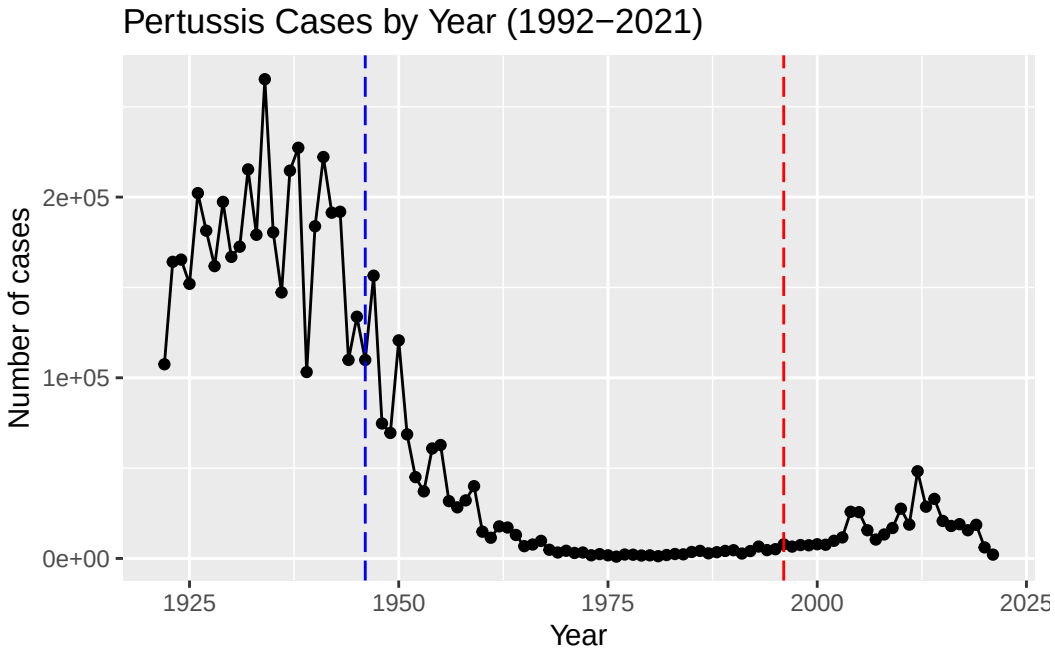
```
library(ggplot2)

casesplot <- ggplot (cdc) +
  aes(x= Year, y= Reported.Cases) +
  geom_point() +
  geom_line() +
  labs(title= "Pertussis Cases by Year (1992-2021)", x="Year", y= "Number of cases")
```

Two Vaccines (wP and 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.

```
casesplot +
  geom_vline(xintercept = 1946, col= "blue", linetype="longdash") +
  geom_vline(xintercept = 1996, col= "red", linetype="longdash")
```



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 number of cases reported stayed low for a few years. However small resurgences of cases were observed after the year 2000. This could be due to improved testing sensitivity, hesitancy from parents leading to less vaccination in children, or less effectiveness from the aP vaccine.

Exploring CMI-PB Data

A systems vaccine project to figure out what is going on with aP vs wP immune responses. Evaluation of pertussis-specific immune responses over time in individuals with wP and aP vaccines. Look into the data from CMI-PB project using jsonlite package, will simplify JSON "key:value" pair arrays into R data frames

```
# load in the jsonlite package
library(jsonlite)

# read in main subject database table from the CMI-PB API
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 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$race, subject$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
```

Working with dates

Converting year-of-birth and date-of-boost columns from dates into years. We use the `lubridate` package in R

> 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?

```
#install lubridate package in R console
library(lubridate)
```

```
Attaching package: 'lubridate'
```

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

    date, intersect, setdiff, union
```

```
#create a column "age" in subject dataframe for the age of all subjects in years
subject$age <- time_length(today() -ymd(subject$year_of_birth), "years")
```

```
#load up dplyr to filter subjects age by aP vs wP
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
```

```
#filter for aP
ap <- subject %>% filter(infancy_vac == "aP")
round(summary(ap$age))
```

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

```
#filter for wP
wp <- subject %>% filter(infancy_vac == "wP")
round(summary(wp$age))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     28      32      36      37      39      56
```

Joining Multiple Tables

Read other tables from the CMi-PB resource

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

```
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_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
```

We must link the specimen and subject data frames to determine whether the specimen_id comes from aP or wP individuals. We shall link the tables into a data frame called "meta"

> Q9. join specimen and subject tables to make a new merged data frame containing all specimen records along with their associated subject details:

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

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

```
dim(meta)
```

```
[1] 939  14
```

```
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 38.40931
2 38.40931
3 38.40931
4 38.40931
5 38.40931
6 38.40931
```

Q10. Now using the same procedure join meta with "ab_titer" data so we can further analyze this data in terms of time of visit aP/wP, male/female etc.

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

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

```
dim(abdata)
```

```
[1] 41775    21
```

```
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
1 38.40931
2 38.40931
3 38.40931
4 38.40931
5 38.40931
6 38.40931
```

Now we have our merged dataset with all the needed metadata and antibody measurements called "abdata"

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 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?

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

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

Examine IgG Ab Titer Levels

Use the dplyr package to again filter our abdata merged dataset by IgG isotype

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

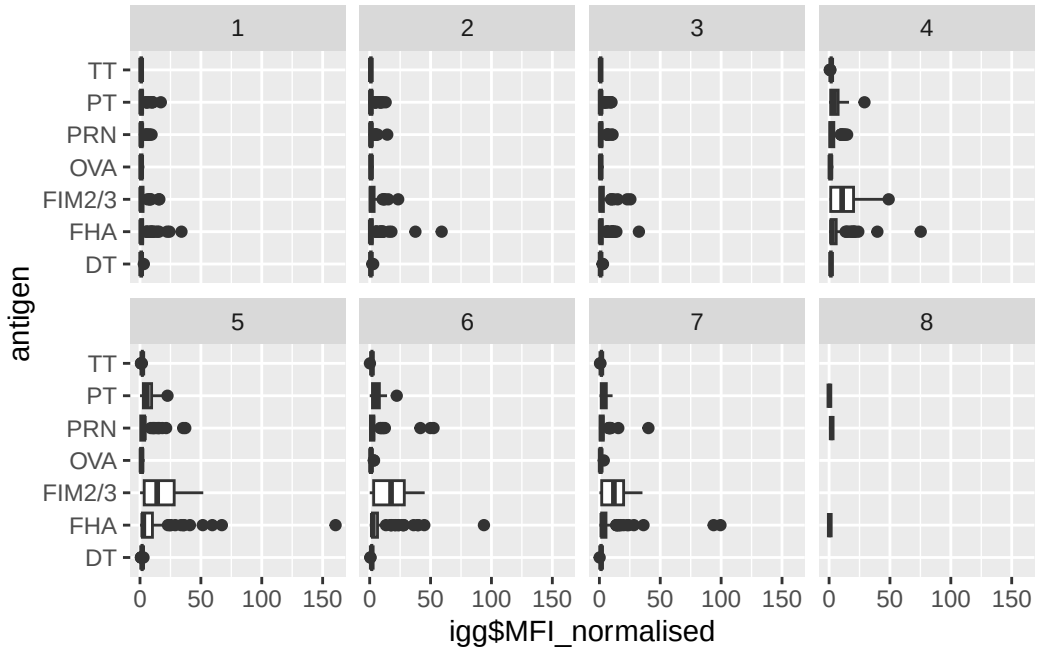
```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgG                TRUE      PT   68.56614       3.736992
2           1     IgG                TRUE     PRN  332.12718       2.602350
3           1     IgG                TRUE     FHA 1887.12263      34.050956
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 38.40931
2 38.40931
3 38.40931
4 41.40999
5 41.40999
6 41.40999
```

Q13. Make a summary boxplot of Ab titer levels (MFI) for all antigens:

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

```
Warning: Use of `igg$MFI_normalised` is discouraged.
i Use `MFI_normalised` instead.
```



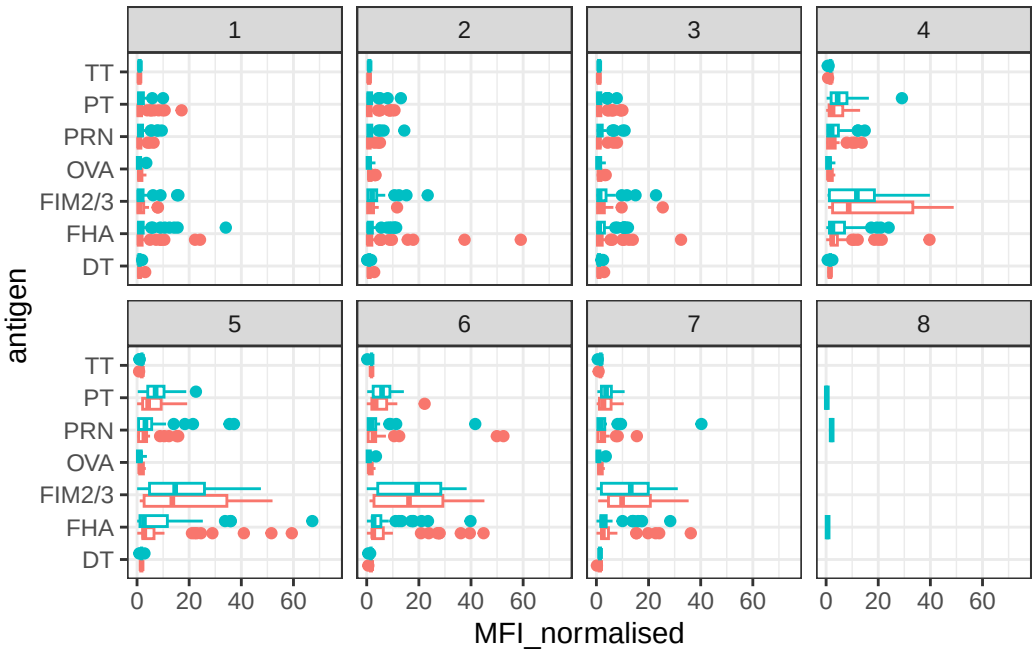

```
table(igg$visit)
```

```
  1   2   3   4   5   6   7   8
524 531 552 426 426 393 378   3
```

Let's dig in a little bit more...

```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

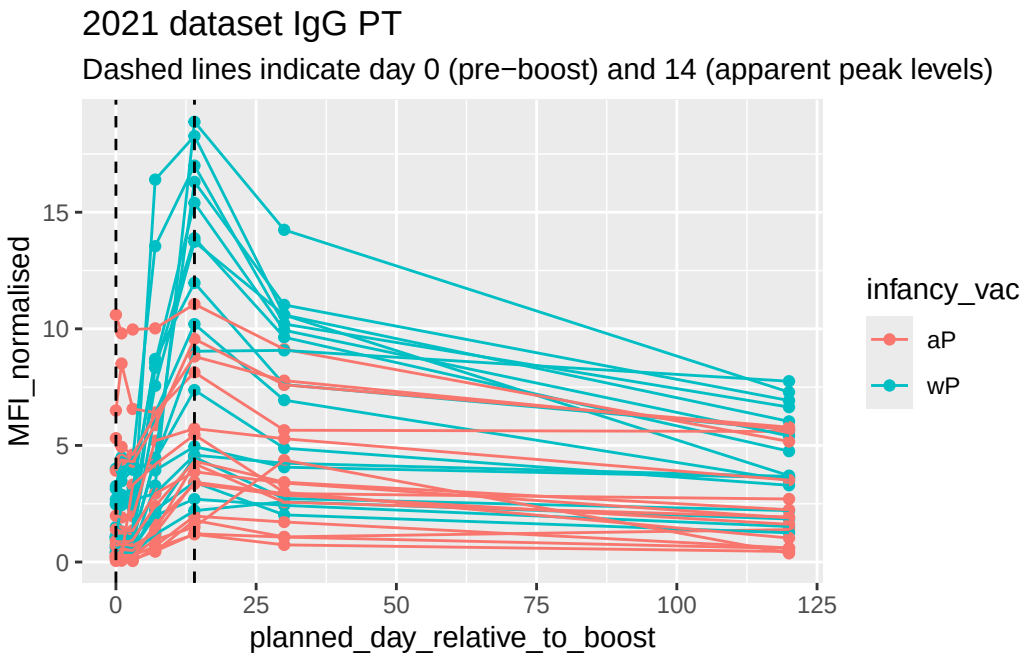
```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



Let's look into the 2021 dataset IgG PT antigen level time-course

```
abdata.21 <- abdata %>% filter(dataset == "2021_dataset")

abdata.21 %>%
  filter(isotype == "IgG",  antigen == "PT") %>%
  ggplot() +
    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)")
```


2021 dataset IgG PT
Dashed lines indicate day 0 (pre−boost) and 14 (apparent peak levels)
MFI_normalised
0
5
10
15
0
25
50
75
100
125
planned_day_relative_to_boost
infancy_vac
aP
wP