Stat 744 Assignment 2

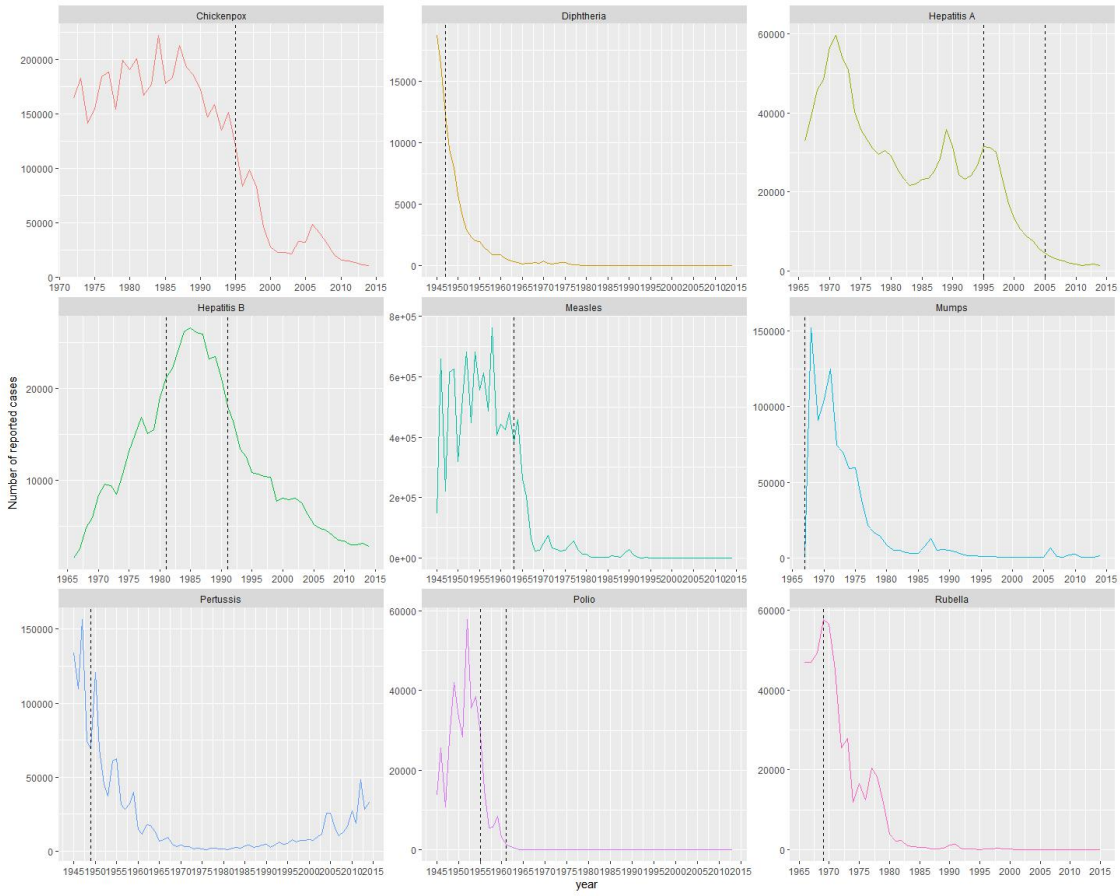
Yusang Hu 400177333

Sep 18, 2019

Jia You's research of interest is studying vaccines for 9 diseases, each disease corresponds to one or more vaccines. She used graphs to show that vaccines have high effectiveness on diseases, the effect lasts a long time and probability of developing resistance to the vaccine is very small. Size of the blue bubble represents the number of people who have disease; yellow circle with dot represents the time when the vaccine was licensed.

First Plot

```
setwd("/Users/admin/Desktop/744")
vaccines=read.csv("vaccine_data_online.csv")
library(ggplot2)
vline=data.frame(disease=c("Diphtheria","Polio","Polio","Measles","Pert
ussis","Hepatitis A","Hepatitis A","Hepatitis B","Hepatitis B","Mumps",
"Rubella","Chickenpox"),year=c(1947,1955,1961,1963,1949,1995,2005,1981,
1991,1967,1969,1995))
(g1 <- ggplot(vaccines,aes(year,cases,colour=disease)) ## x=year, y=cas
es, each color represents a disease
  + geom_line()
  + scale_x_continuous(breaks=seq(1945,2015,by=5)) ## rescale x-axis
  + labs(y="Number of reported cases")
  + facet_wrap(~disease,scales="free") ## set 'free' scale to make each
facet looks better
  + theme(legend.position="none")
  + geom_vline(aes(xintercept=year),vline,linetype="dashed") ## dashed
line represents the time of vaccine was licensed
)
```



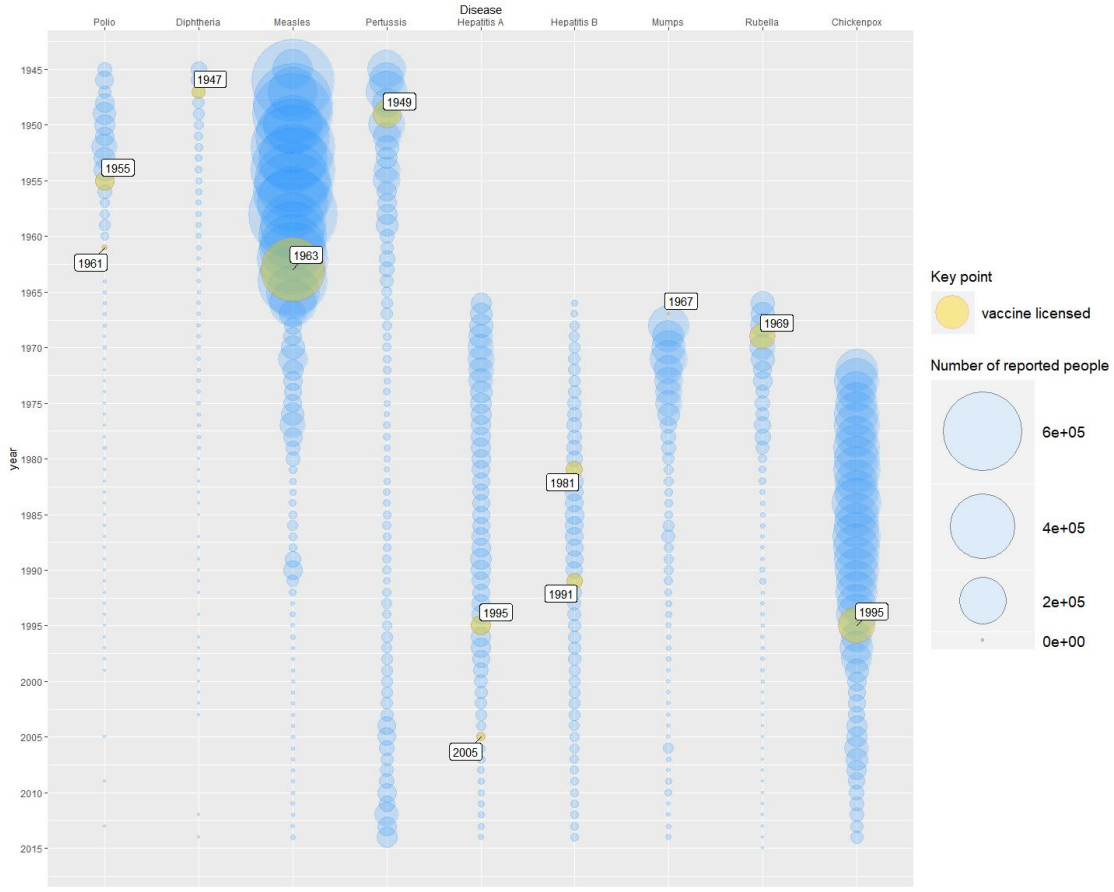
Comments on First Plot & Comparing with the Original Plot

I choose to use facet wrap plot as my first plot because each disease is independent to others and corresponds to one or more vaccines, therefore results will be more clearer if I separated 9 diseases and create graphs for each of them individually. I set x-axis as year and y-axis as number of people who have disease, each disease has its own color. In facet_wrap, I set scale as 'free' because this 9 diseases have different origin (i.e. Diphtheria was first reported on 1945, Hepatitis B was first reported on 1966) and different population proportion of disease (i.e. population proportion of Diphtheria is small, population proportion of Measles is large). Free scale will set the most suitable scale automatically, which will make each facet looks better, clearer and more easier to interpret. Also, I used dashed line to represent the time when the vaccine was licensed. This will be easier to compre the number of people who have disease before and after the vaccine was licensed.

Advantages of my first plot are: it is easy to see trends within each facet and differences between the number of people who have disease before and after the vaccine was licensed. Disadvantages of my first plot are: some diseases have two kinds of vaccines, for example Hepatitis A, it is hard to say which vaccine has better effectiveness on Hepatitis A, and also facet wrap plot is hard to see overall relationship between each facet.

My first plot answer the question better than the original plot because I solved the problem of overlapping of bubbles. In the original plot, a lot of bubbles are mixed together, so it's hard to see what the trend exactly looks like. My first plot has very clear trends for each disease, it is easy to identify when the number of people who have disease decreases or increases, therefore it's easy to interpret some problems such as how the vaccine works, is it possible to develop resistance to the vaccine.

Second Plot



Comments on Second Plot & Comparing with the Original Plot

In my second plot, I used the same graphical design with original plot but with some modifications. Same with the original plot, I set x-axis as disease and y-axis as year, but I reordered diseases (x-axis) by the time they were first reported. Size of bubbles are depended on the number of people who have disease, and I set alpha=0.2 to make bubbles transparent which will somewhat avoid overlappings. Gold bubbles represent the time when the vaccine was licensed. In this plot, all diseases are in one big plot, so in order to make it easier to identify the vaccine licensing time, I made a label next to each of gold bubbles which represents the exact vaccine licensing time.

Advantage of my second plot is: I reordered diseases by year, so my second plots looks better than the original plot. Disadvantages of my second plot and the orginal plot are: first, the problem of overlapping bubbles, we are only able to see the overall trend, but it is hard to see small details within trend; it is not a good idea to put all 9 diseases in one plot since they have different affected population proportion, each disease is supposed to have their own scale. For example, difference between the number of people who got Measles from 1945 to the vaccine was licensed (1945s to 1963s) and from the start of vaccine licensing to 2015 (1963s to 2015s) is very large, so it is easy to interpret that the vaccine has high effectiveness; but for Heptitis B, it is hard to say that whether the vaccine works well or not since Heptitis B's original proportion of affected population is small, or say more techinically, the visual differences for Heptitis B is too small.