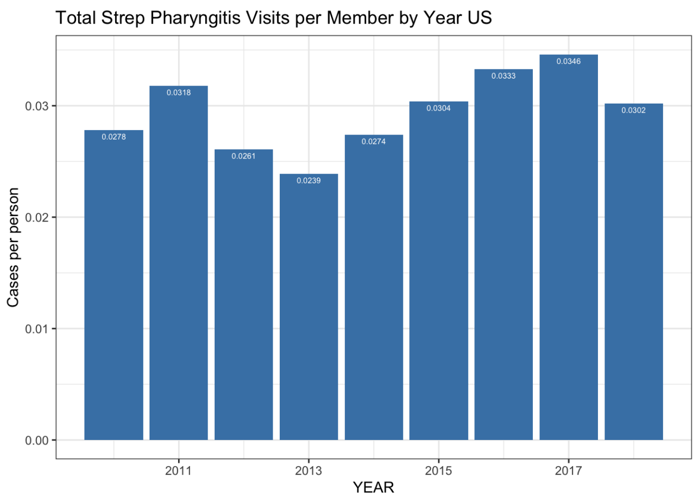
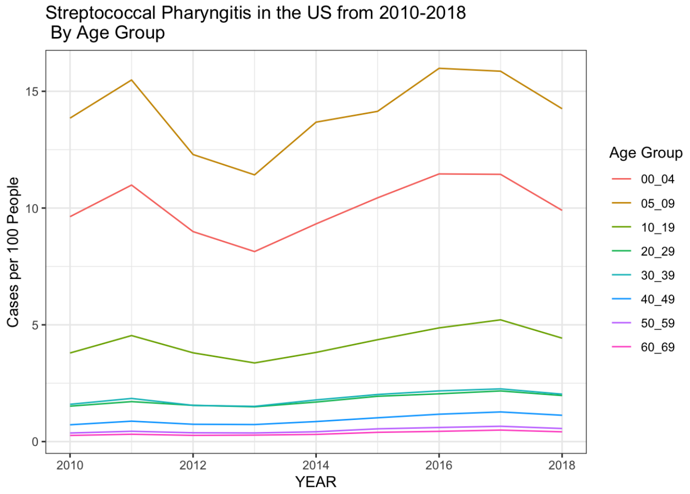
Introduction

Streptococcal pharyngitis, known colloquially as “strep throat,” is the number one cause of pharyngitis in children, accounting for 15-30% of pharyngitis between the ages of 5-15, and up to as many as 35-40% of cases during the winter and early spring. There is an estimated prevalence of 37% among children[1][2][3] . Symptoms typically include a characteristic sore throat, sometimes alongside fever, abdominal pain, or headache. Streptococcal pharyngitis is caused by group A *Streptococcus*, or *Streptococcus pyogenes*, and is clearly diagnosed with the rapid antigen detection test, throat culture, or NAAT. Treatment with the appropriate penicillin antibiotics is recommended to decrease the chance of some complications, including acute rheumatic fever, which can cause lasting cardiac valvular damage. There are known geographic patterns across the US for some other respiratory illnesses, including influenza, which starts in the Southeast and subsequently rises in other parts of the country [4]. Comparatively little is known about the geographic distribution of streptococcal pharyngitis in the US.

The MarketScan database, comprised of insurance claims data from private insurers across the US, provides an opportunity to investigate the distribution of streptococcal pharyngitis in the US. Claims data for streptococcal pharyngitis were extracted from MarketScan, separated by year from 2010 to 2018, month, sex, age group (by decile with the youngest age group split between 0-4 and 5-9), and state. Data for flu with the same properties were also extracted from MarketScan for comparison, and vaccination data for flu and PCV were obtained through CDC FluVax and ChildVax respectively. The question we seek to address is whether rates of streptococcal pharyngitis are different in different parts of the country throughout the year and across years, and if so, what factors contribute to these differences.

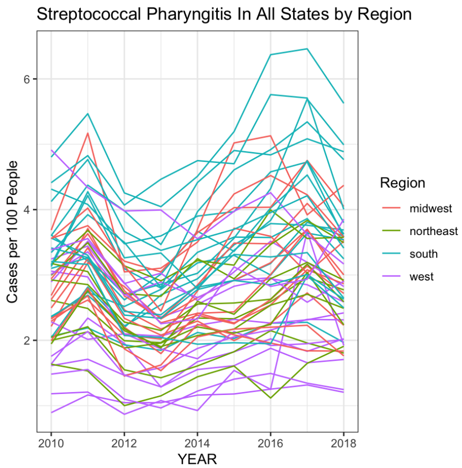


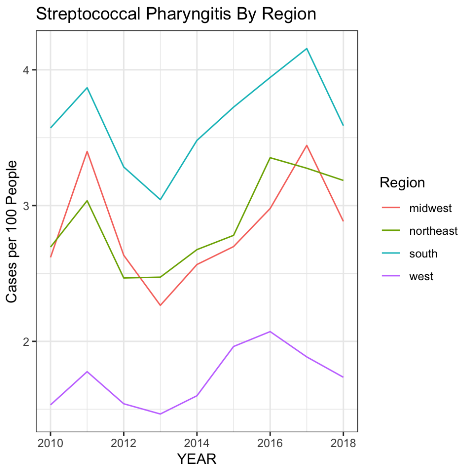
We can see from this graph that the cases per person of streptococcal pharyngitis are relatively stable from year to year between 2010-2018, with slightly more cases in 2011 and 2016-217. We next separate by age.



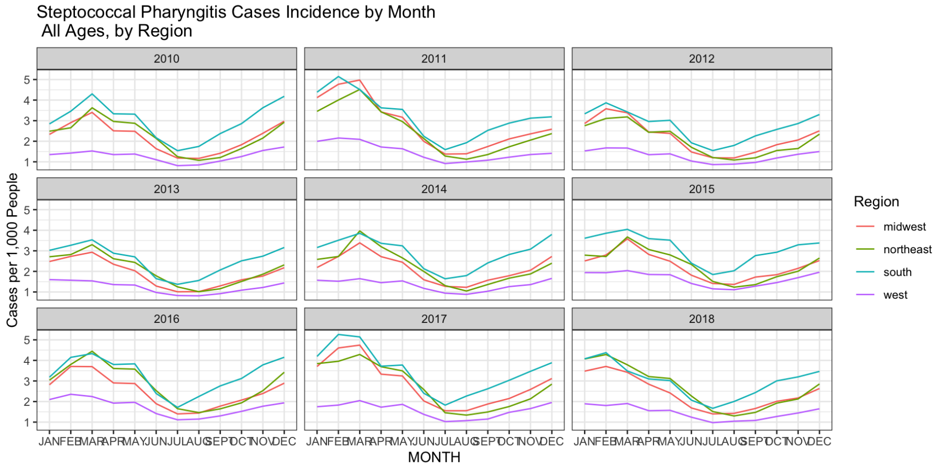
Here we see that streptococcal pharyngitis is most common in 5-9 year olds, followed by 0-4 year olds, then 10-19 year olds, and there are not many cases in people older than 20, which is consistent with the known age distribution of streptococcal pharyngitis as stated above.

We can split this data by region to see if there are trends that differ in different parts of the country. We divide the country into 4 regions, Northeast (comprised of Connecticut, Massachusetts, Maine, New Hampshire, Rhode Island, Vermont, New Jersey, New York, and Pennsylvania), Midwest (comprised of Illinois, Indiana, Michigan, Ohio, Wisconsin, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakata), South (comprised of Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia, Alabama, Kentucky, Mississippi, Tennessee, Arizona, Louisiana, Oklahoma, and Texas), and West (comprised of Arizona, Colorado, Indiana, Montana, Nevada, New Mexico, Utah, Wyoming, Arkansas, California, Hawaii, Oregon, and Washington).





From this plot we can see that streptococcal pharyngitis cases are generally higher in the South, and lower in the West, and that this trend holds across time. We can further stratify case data by month to see if there are differences in trends throughout the year by region.



We see that the West has fewer cases of streptococcal pharyngitis throughout the year. The South tracks pretty well with the Northeast and Midwest through the beginning of the year, but cases rise earlier and higher starting at the end of the summer and into the fall. My analysis will seek to identify what factors drive this regional difference in case rates.

Appendix

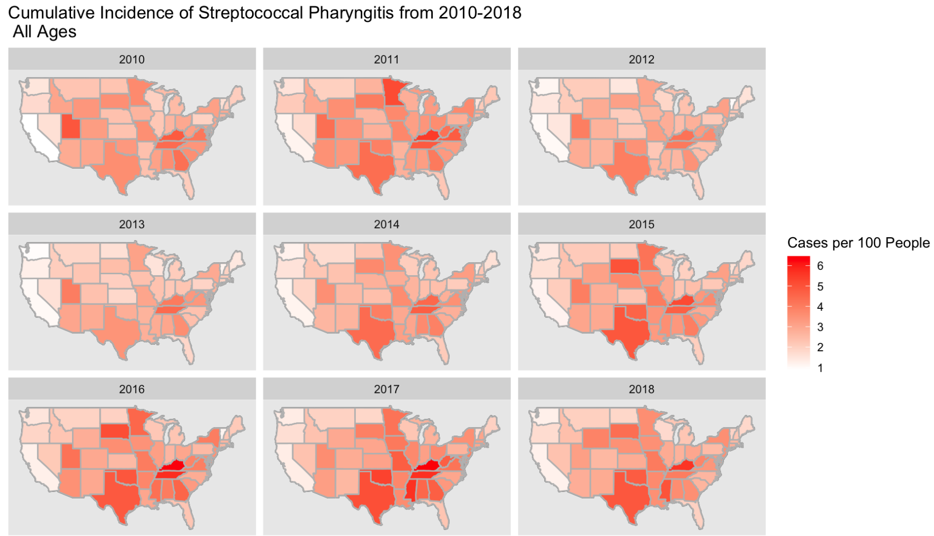
```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(stringr)
library(ggplot2)
library(usmap)
library(gridExtra)
library(maps)
library(mapdata)
library(ggmap)
library(lmtest)

#Read in data on cases
dat <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/StrepPharyngitis/output/Ge
oVisits.csv")

#Read in data on membership
coh <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/StrepPharyngitis/output/Ge
oCohort.csv")
#this dataframe has population by sex, age group, and state but not by month. So need to add it to the ot
her dataframe once already collapsed by year.

#get data by year, rather than by month:
by_year <- aggregate(NVISITS ~ YEAR + STATE + SEX + AGEGRP + PRIMARYCOND, dat, sum)
#we can then join this with the population data from cohort
by_year <- left_join(by_year, coh)

by_year |> group_by(YEAR) |> summarize(total_vis = sum(NVISITS), total_memb = sum(NMEMB), CI = round(tota
l_vis/total_memb,4)) |> ggplot(aes(x=YEAR, y = CI)) + geom_bar(stat = "identity", fill = "steelblue") + t
heme_bw() + geom_text(aes(label=CI), vjust=1.6, color="white", size=2.0) + ylab("Cases per person") +  gg
title("Total Strep Pharyngitis Visits per Member by Year US")


#let's summarize cumulative incidence by state across years to start
by_state <- by_year |>
  group_by(YEAR, STATE) |>
  summarize(visits = sum(NVISITS), members = sum(NMEMB), CI = visits/members)

#we now manipulate the data slightly to show CI per hundred, and make the state names match the mapping d
ataframe
strep_all <- by_state |> mutate(region = tolower(STATE), CI_per_hundred = CI*100) |>
  select(region, CI_per_hundred)



#now we add region designations for "northeast", "south", "midwest", and "west"
#add regions to this dataframe
#make a function that converts lists of state abbreviations to lists of state names
to_statename <- function(list){
  new_list <- c()
  for(i in 1:length(list)){
    name <- state.name[grep(list[i], state.abb)]
    new_list <- append(new_list, name)
  }
  new_list
}

northeast_states <- tolower(to_statename(.northeast_region))
midwest_states <- tolower(to_statename(.midwest_region))
south_states <- tolower(to_statename(.south_region))
west_states <- tolower(to_statename(.west_region))

northeast_df <- data.frame(region = northeast_states, part = "northeast")
midwest_df <- data.frame(region = midwest_states, part = "midwest")
south_df <- data.frame(region = south_states, part = "south")
west_df <- data.frame(region = west_states, part = "west")
#will put dc in the south because maryland and viriginia are
```

```
dc_df <- data.frame(region = "washington dc", part = "south")


state_parts <- rbind(northeast_df, midwest_df, south_df, west_df, dc_df)


strep_all_region <- left_join(strep_all, state_parts, by = "region")

#now we look at the data by age group across all regions
by_year <- by_year |> mutate("state" = STATE)
by_year_age_visits <- aggregate(NVISITS ~ YEAR + state + AGEGRP + PRIMARYCOND, dat = by_year, sum)
by_year_age_membs <- aggregate(NMEMB ~ YEAR + state + AGEGRP + PRIMARYCOND, dat = by_year, sum)
by_year_age <- left_join(by_year_age_visits, by_year_age_membs)
by_year_age <- by_year_age |> mutate(CI_per_hundred = NVISITS/NMEMB *100, state = tolower(state))

#add in region just in case?
state_parts_2 <- state_parts
names(state_parts_2)[1] <- "state"

by_year_age <- left_join(by_year_age, state_parts_2)
#make a plot of trends over time by state


country_by_age <- left_join(aggregate(NVISITS ~ AGEGRP + YEAR, dat = by_year_age, sum), aggregate(NMEMB ~
AGEGRP + YEAR, dat = by_year_age, sum))
country_by_age <- country_by_age |> mutate(CI_per_hundred = NVISITS/NMEMB * 100)

country_by_age |> ggplot(aes(YEAR, CI_per_hundred, group = AGEGRP)) + geom_line(aes(color = AGEGRP)) + gg
title("Streptococcal Pharyngitis in the US from 2010-2018 \n By Age Group") + theme_bw() + ylab("Cases pe
r 100 People") + labs(color = "Age Group")



states_indiv_region <- strep_all_region |> group_by(YEAR, region) |> ggplot(aes(YEAR, CI_per_hundred, gro
up=region)) + geom_line(aes(col = part)) + ggtitle("Streptococcal Pharyngitis In All States by Region") +
ylab("Cases per 100 People") + theme_bw() + labs(color = "Region")

#now group them by region and just report 1 value per region

by_state_2 <- by_state |> mutate(state = tolower(STATE))
strep_region_visits <- left_join(by_state_2, state_parts_2, by = "state")
strep_region_visits_agg <- aggregate(visits~ part + YEAR, dat = strep_region_visits, sum)
strep_region_members_agg <- aggregate(members~ part + YEAR, dat = strep_region_visits, sum)
strep_region_joined <- left_join(strep_region_visits_agg, strep_region_members_agg)
strep_region_joined <- strep_region_joined |> mutate(CI_per_hundred = visits/members * 100)

per_region<- strep_region_joined |> group_by(YEAR, part) |> ggplot(aes(YEAR, CI_per_hundred, group=part))
+ geom_line(aes(col = part)) + ggtitle("Streptococcal Pharyngitis By Region") +ylab("Cases per 100 Peopl
e") + theme_bw() + labs(color = "Region")

grid.arrange(states_indiv_region, per_region, ncol=2)

#try by month; first just plot data by region for all age groups together
#should be able to just merge these two because the members are stable over the year
by_month_all <- left_join(dat, coh)

#consolidate across sex and check if this is what you needed to do earlier on as well;
#to consolidate across sex, will just need to add so that should be fine
by_month_age_vis <- aggregate(NVISITS ~ MONTH + STATE + AGEGRP + YEAR, dat = by_month_all, sum)
memb_no_sex <- aggregate(NMEMB ~  STATE +AGEGRP + YEAR, dat = coh, sum)
by_month_no_sex <- left_join(by_month_age_vis, memb_no_sex)
by_month_no_sex <- by_month_no_sex |> mutate("CI_per_thousand" = NVISITS/NMEMB*1000)



by_month_no_sex_lowercase <- by_month_no_sex |> mutate("state" = tolower(STATE))
by_month_regions <- left_join(by_month_no_sex_lowercase, state_parts_2, by = "state")
by_month_regions_only_vis <- aggregate(NVISITS ~ part  + YEAR + MONTH,
                                       dat = by_month_regions,
                                       sum)
by_month_regions_only_memb <- aggregate(NMEMB ~ part  + YEAR + MONTH,
                                       dat = by_month_regions,
```

```
                                        sum)


by_month_regions_only <- left_join(by_month_regions_only_vis, by_month_regions_only_memb)
by_month_regions_only <- by_month_regions_only |> mutate("CI_per_thousand" = NVISITS/NMEMB *1000)
by_month_regions_only |> ggplot(aes(x = MONTH, y = CI_per_thousand, group = part)) +
  geom_line(aes(color = part)) +
  facet_wrap(~YEAR) + ggtitle("Steptococcal Pharyngitis Cases Incidence by Month \n All Ages, by Region")
+ labs(color = "Region") + ylab("Cases per 1,000 People") +scale_x_discrete(limits = c("JAN", "FEB", "MA
R", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT",

"OCT", "NOV", "DEC")) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + theme_bw()
```

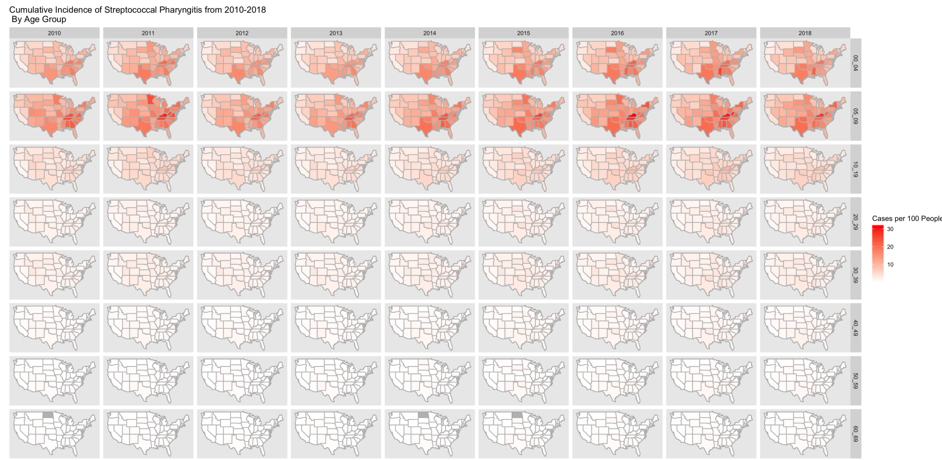
Mapping

As a first step in my analysis, I will do a descriptive analysis looking at case rates of streptococcal pharyngitis across states in the US over time.



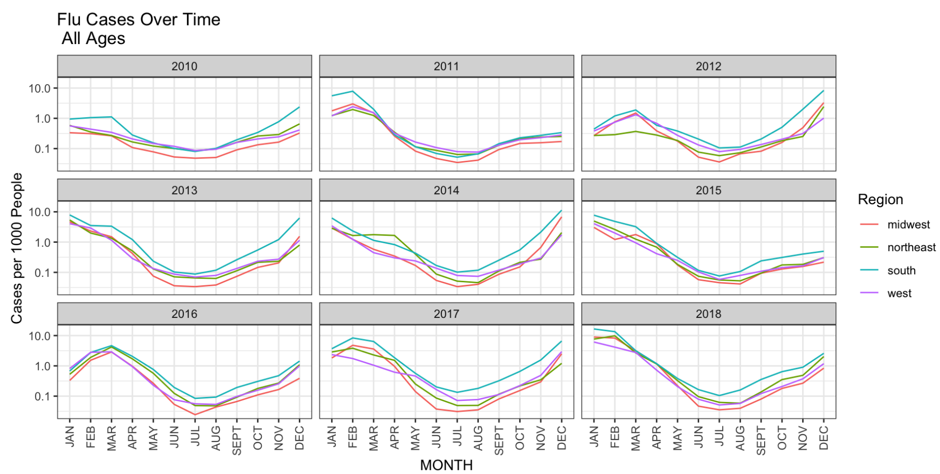
From this plot, we can see that that, across all age groups, cases of streptococcal pharyngitis are much higher in states like Texas, Oklahoma, Mississippi, Tennessee and Kentucky, especially in more recent years (2016, 2017, 2018), and states in the Pacific Northwest such as California, Oregon, Washington, Montana and North Dakota, have lower case rates.

We can look at how cases trend by age group over time.



From these plots, we can see that the cases of streptococcal pharyngitis are highest in the 5-9 year old age group, as previously explained, and that there are relatively few cases in the older age groups, but cases persist at a low level across the country in the older age groups.

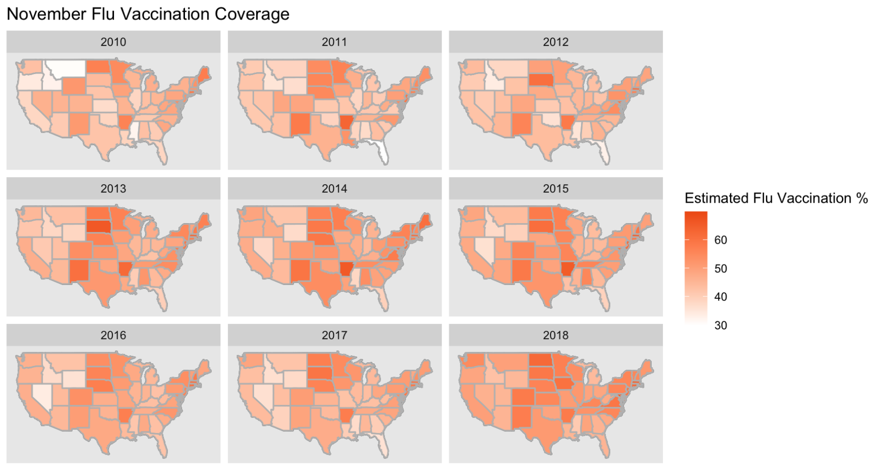
How closely do these trends track with flu data? We can obtain the same data from MarketScan for flu data, using ICD codes: ICD9 4870, 4871, 7878, ICD10: J100, J101, J108, J110, J111, and J118.



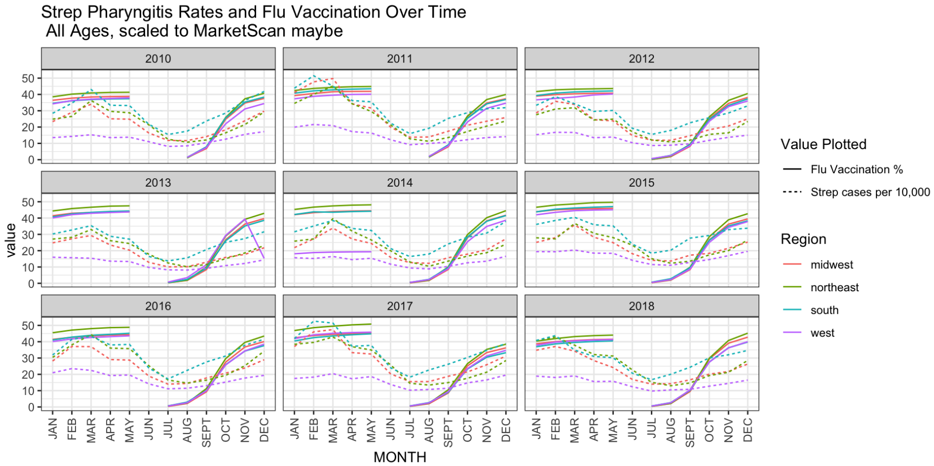
Here, we see similar regional trends play out, with a similar early fall divergence in the South compared to the other states. One notable difference is that the West region does not show a similar lack of disease burden; it follows a similar pattern as the Midwest and Northeast regions. This is important because it indicates that the regional trend observed for streptococcal pharyngitis cases is not entirely explained by factors such as selection bias, whereby the privately insured people in the West may be more likely to be wealthier and healthier because these states have a more robust public healthcare system, compared to states in the South with relatively limited public healthcare system, where a more representative fraction of the population may be covered by public health insurance.

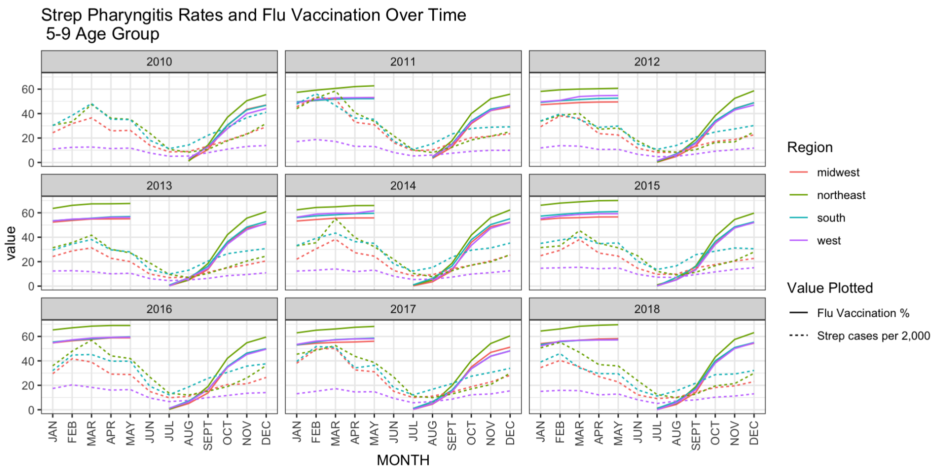
We could consider a situation in which people who have flu are more likely to get a secondary bacterial infection with streptococcus, or that the factors driving an increase in flu in the South are the same as the factors driving the increase in streptococcal pharyngitis. Will will consider flu cases in our statistical model of streptococcal pharyngitis cases.

If flu drives cases of streptococcal pharyngitis, it would also be useful to look at how trends in flu vaccination map to cases of streptococcal pharyngitis. We obtain flu vaccination from the CDC website FluVaxView [5].



Here we see that flu vaccination rates are generally not especially high; ranging from 30-70% by November. Some states, especially in the Midwest, seem to have consistently higher flu vaccination than others, like Florida.

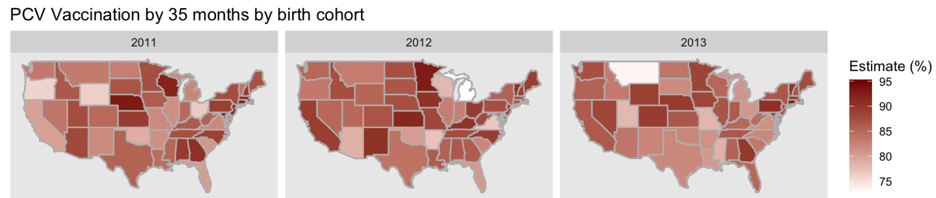




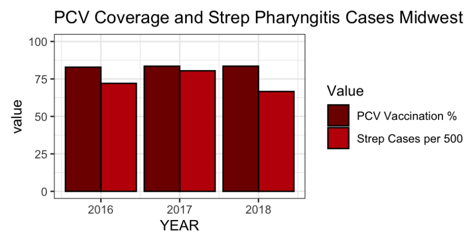
Data on flu vaccination starts in a given year in July, and is collected through May, then restarts in July again. We can compare flu vaccination coverage to streptococcal pharyngitis cases in all ages vs in just the 5-9 age group, which is the highest prevalence age group, as seen previously. While in general, the Northeast has high rates of vaccination, and the other regions cluster pretty closely, there are a few examples of flu vaccination being lower in the South, particularly in the 2016-2017 and 2017-2018 flu seasons in all ages, although this does not seem to be the case in the 5-9 age group, where overall vaccination rates are generally higher.

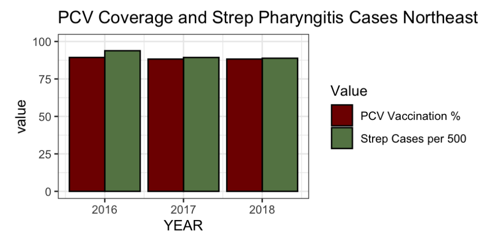
We are also interested in whether vaccination with PCV is protective against streptococcal pharyngitis.Prior research has shown that introduction of current PCV vaccines corresponded with a decline in overall cases of pharyngitis (including cases caused by pathogens other than Group A *Streptococcus*)[6]. There is also some evidence that the PCV vaccine (against *Streptococcus pneumoniae*) can prevent early cases of otitis media, infection of the middle ear that can be caused by pneumococcus or other pathogens [7][8].

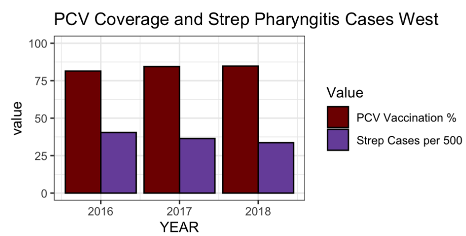
Data on PCV vaccination were obtained from CDC ChildVaxView [9]. The CDC currently recommends that children receive 4 doses of either PCV13 or PCV15 at 2 months, 4 months, 6 months, and 12-15 months old. Children who miss doses are recommended to follow catch-up guidelines to receive their vaccinations [9]. The CDC reports this vaccination data by birth cohort from 2011-2018. Here we use data on whether children received all 4 PCV vaccinations by the age of 35 months. We use the birth cohorts from 2011, 2012, and 2013 to compare to flu and streptococcal pharyngitis cases occurring in 2016, 2017, and 2018 in those in the 5-9 age group.

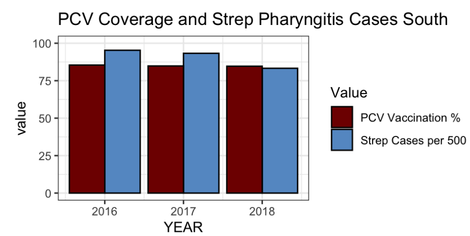


These data show generally high vaccination (>75% universally), with some variation from year to year as to which states have the highest coverage. We can group these birth cohorts to plot onto our case count data from streptococcal pharyngitis and flu by including the 2011 birth cohort vaccination status for 5-9 year olds in the year 2016, 2011 and 2012 for 5-9 year olds in 2017, and 2011, 2012, and 2013 for 2018.









PCV coverage is high across the regions in these years, and no clear trends jump out of this visualization. However, there are some variations from year to year in both cases and vaccination. For example, PCV vaccination appears to increase in the West over these 3 years, and cases of streptococcal pharyngitis go down.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(stringr)
library(ggplot2)
library(usmap)
library(gridExtra)
library(maps)
library(mapdata)
library(ggmap)
library(lmtest)

#Read in data on cases
dat <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/StrepPharyngitis/output/Ge
oVisits.csv")

#Read in data on membership
coh <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/StrepPharyngitis/output/Ge
oCohort.csv")
#this dataframe has population by sex, age group, and state but not by month. So need to add it to the ot
her dataframe once already collapsed by year.

#get data by year, rather than by month:
by_year <- aggregate(NVISITS ~ YEAR + STATE + SEX + AGEGRP + PRIMARYCOND, dat, sum)
#we can then join this with the population data from cohort
by_year <- left_join(by_year, coh)

# by_year |> group_by(YEAR) |> summarize(total_vis = sum(NVISITS), total_memb = sum(NMEMB), CI = round(to
tal_vis/total_memb,4)) |> ggplot(aes(x=YEAR, y = CI)) + geom_bar(stat = "identity", fill = "steelblue") +
theme_bw() + geom_text(aes(label=CI), vjust=1.6, color="white", size=2.0) + ylab("Cases per person") +  g
gtitle("Total Strep Pharyngitis Visits per Member by Year US")


#let's summarize cumulative incidence by state across years to start
by_state <- by_year |>
  group_by(YEAR, STATE) |>
  summarize(visits = sum(NVISITS), members = sum(NMEMB), CI = visits/members)

#we now manipulate the data slightly to show CI per hundred, and make the state names match the mapping d
ataframe
strep_all <- by_state |> mutate(region = tolower(STATE), CI_per_hundred = CI*100) |>
  select(region, CI_per_hundred)



#now we add region designations for "northeast", "south", "midwest", and "west"
#add regions to this dataframe
#make a function that converts lists of state abbreviations to lists of state names
to_statename <- function(list){
  new_list <- c()
  for(i in 1:length(list)){
    name <- state.name[grep(list[i], state.abb)]
    new_list <- append(new_list, name)
  }
  new_list
}

northeast_states <- tolower(to_statename(.northeast_region))
midwest_states <- tolower(to_statename(.midwest_region))
south_states <- tolower(to_statename(.south_region))
west_states <- tolower(to_statename(.west_region))

northeast_df <- data.frame(region = northeast_states, part = "northeast")
midwest_df <- data.frame(region = midwest_states, part = "midwest")
south_df <- data.frame(region = south_states, part = "south")
west_df <- data.frame(region = west_states, part = "west")
#will put dc in the south because maryland and viriginia are
```

```
dc_df <- data.frame(region = "washington dc", part = "south")


state_parts <- rbind(northeast_df, midwest_df, south_df, west_df, dc_df)


strep_all_region <- left_join(strep_all, state_parts, by = "region")

#now we look at the data by age group across all regions
by_year <- by_year |> mutate("state" = STATE)
by_year_age_visits <- aggregate(NVISITS ~ YEAR + state + AGEGRP + PRIMARYCOND, dat = by_year, sum)
by_year_age_membs <- aggregate(NMEMB ~ YEAR + state + AGEGRP + PRIMARYCOND, dat = by_year, sum)
by_year_age <- left_join(by_year_age_visits, by_year_age_membs)
by_year_age <- by_year_age |> mutate(CI_per_hundred = NVISITS/NMEMB *100, state = tolower(state))

#add in region just in case?
state_parts_2 <- state_parts
names(state_parts_2)[1] <- "state"

by_year_age <- left_join(by_year_age, state_parts_2)
#make a plot of trends over time by state


country_by_age <- left_join(aggregate(NVISITS ~ AGEGRP + YEAR, dat = by_year_age, sum), aggregate(NMEMB ~
AGEGRP + YEAR, dat = by_year_age, sum))
country_by_age <- country_by_age |> mutate(CI_per_hundred = NVISITS/NMEMB * 100)

# country_by_age |> ggplot(aes(YEAR, CI_per_hundred, group = AGEGRP)) + geom_line(aes(color = AGEGRP)) +
ggtitle("Streptococcal Pharyngitis in the US from 2010-2018 \n By Age Group") + theme_bw() + ylab("Cases
per 100 People") + labs(color = "Age Group")



states_indiv_region <- strep_all_region |> group_by(YEAR, region) |> ggplot(aes(YEAR, CI_per_hundred, gro
up=region)) + geom_line(aes(col = part)) + ggtitle("Streptococcal Pharyngitis In All States by Region") +
ylab("Cases per 100 People") + theme_bw() + labs(color = "Region")

#now group them by region and just report 1 value per region

by_state_2 <- by_state |> mutate(state = tolower(STATE))
strep_region_visits <- left_join(by_state_2, state_parts_2, by = "state")
strep_region_visits_agg <- aggregate(visits~ part + YEAR, dat = strep_region_visits, sum)
strep_region_members_agg <- aggregate(members~ part + YEAR, dat = strep_region_visits, sum)
strep_region_joined <- left_join(strep_region_visits_agg, strep_region_members_agg)
strep_region_joined <- strep_region_joined |> mutate(CI_per_hundred = visits/members * 100)

per_region<- strep_region_joined |> group_by(YEAR, part) |> ggplot(aes(YEAR, CI_per_hundred, group=part))
+ geom_line(aes(col = part)) + ggtitle("Streptococcal Pharyngitis By Region") +ylab("Cases per 100 Peopl
e") + theme_bw() + labs(color = "Region")

#grid.arrange(states_indiv_region, per_region, ncol=2)

#try by month; first just plot data by region for all age groups together
#should be able to just merge these two because the members are stable over the year
by_month_all <- left_join(dat, coh)

#consolidate across sex and check if this is what you needed to do earlier on as well;
#to consolidate across sex, will just need to add so that should be fine
by_month_age_vis <- aggregate(NVISITS ~ MONTH + STATE + AGEGRP + YEAR, dat = by_month_all, sum)
memb_no_sex <- aggregate(NMEMB ~  STATE +AGEGRP + YEAR, dat = coh, sum)
by_month_no_sex <- left_join(by_month_age_vis, memb_no_sex)
by_month_no_sex <- by_month_no_sex |> mutate("CI_per_thousand" = NVISITS/NMEMB*1000)



by_month_no_sex_lowercase <- by_month_no_sex |> mutate("state" = tolower(STATE))
by_month_regions <- left_join(by_month_no_sex_lowercase, state_parts_2, by = "state")
by_month_regions_only_vis <- aggregate(NVISITS ~ part  + YEAR + MONTH,
                                       dat = by_month_regions,
                                       sum)
by_month_regions_only_memb <- aggregate(NMEMB ~ part  + YEAR + MONTH,
                                       dat = by_month_regions,
```

```
                        sum)


by_month_regions_only <- left_join(by_month_regions_only_vis, by_month_regions_only_memb)
by_month_regions_only <- by_month_regions_only |> mutate("CI_per_thousand" = NVISITS/NMEMB *1000)
# by_month_regions_only |> ggplot(aes(x = MONTH, y = CI_per_thousand, group = part)) +
#   geom_line(aes(color = part)) +
#   facet_wrap(~YEAR) + ggtitle("Steptococcal Pharyngitis Cases Incidence by Month \n All Ages, by Regio
n") + labs(color = "Region") + ylab("Cases per 1,000 People") +scale_x_discrete(limits = c("JAN", "FEB",
"MAR", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT",
#
"OCT", "NOV", "DEC")) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))



#we now load in US mapping data from the maps package
us_df <- map_data("state")

#this creates a base US map to build upon
us_base <- ggplot(data = us_df, mapping = aes(x = long, y = lat, group =group))+
  coord_fixed(1.3) + geom_polygon(color="black", fill = "gray")

#we join our dataframe with strep info with our mapping dataframe
us_strep_all <- inner_join(us_df, strep_all, by = "region")
us_strep_rates_over_time <- us_base + geom_polygon(data = us_strep_all, aes(fill = CI_per_hundred))+
  geom_polygon(color = "gray", fill = NA) + theme(
  axis.text = element_blank(),
  axis.line = element_blank(),
  axis.ticks = element_blank(),
  panel.border = element_blank(),
  panel.grid = element_blank(),
  axis.title = element_blank()
  ) + facet_wrap(~YEAR) + scale_fill_gradient(low = "#FFFFFF", high = "#FF0000" ) + ggtitle("Cumulative I
ncidence of Streptococcal Pharyngitis from 2010-2018 \n All Ages") + labs(fill = "Cases per 100 People")
us_strep_rates_over_time


#making maps for age group trends over time
by_year_age_formap <- by_year_age
names(by_year_age_formap)[2] = "region"
us_strep_byage_map <- inner_join(us_df, by_year_age_formap, by = "region")
#attempt facet_grid with year as the column and age group as the row
us_strep_byage_map_gg <- us_base +
  geom_polygon(data = us_strep_byage_map, aes(fill = CI_per_hundred)) +
    geom_polygon(color = "gray", fill = NA) +
  theme(
  axis.text = element_blank(),
  axis.line = element_blank(),
  axis.ticks = element_blank(),
  panel.border = element_blank(),
  panel.grid = element_blank(),
  axis.title = element_blank()
  ) + scale_fill_gradient(low = "#FFFFFF", high = "#FF0000" ) + ggtitle("Cumulative Incidence of Streptoc
occal Pharyngitis from 2010-2018 \n By Age Group") + labs(fill = "Cases per 100 People")

us_strep_byage_map_gg + facet_grid(rows = vars(AGEGRP), cols = vars(YEAR) )
#now for flu case data
flu_dat <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/StrepPharyngitis/outpu
t/GeoVisitsFlu.CSV")
flu_coh <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/StrepPharyngitis/outpu
t/GeoCohortFlu.csv") #this is the same as strep data

#do same thing that did for strep cases to start
#should be able to just merge these two because the members are stable over the year
by_month_all_flu <- left_join(flu_dat, flu_coh)

flu_by_month_age_vis <- aggregate(NVISITS ~ MONTH + STATE + AGEGRP + YEAR, dat = by_month_all_flu, sum)
flu_memb_no_sex <- aggregate(NMEMB ~  STATE +AGEGRP + YEAR, dat = flu_coh, sum)
flu_by_month_no_sex <- left_join(flu_by_month_age_vis, flu_memb_no_sex)
flu_by_month_no_sex <- flu_by_month_no_sex |>
  mutate("CI_per_thousand" = NVISITS/NMEMB*1000) |>
  mutate("CI_per_ten_thousand" =NVISITS/NMEMB*10000 )
```

```
flu_by_month_no_sex_lowercase <- flu_by_month_no_sex |> mutate("state" = tolower(STATE))
flu_by_month_regions <- left_join(flu_by_month_no_sex_lowercase, state_parts_2, by = "state")
flu_by_month_regions_only_vis <- aggregate(NVISITS ~ part  + YEAR + MONTH,
                                       dat = flu_by_month_regions,
                                       sum)
flu_by_month_regions_only_memb <- aggregate(NMEMB ~ part  + YEAR + MONTH,
                                       dat = flu_by_month_regions,
                                       sum)

flu_by_month_regions_only <- left_join(flu_by_month_regions_only_vis, flu_by_month_regions_only_memb)
flu_by_month_regions_only <- flu_by_month_regions_only |> mutate("CI_per_thousand" = NVISITS/NMEMB *1000,
                                                                 "CI_per_ten_thousand" = NVISITS/NMEMB *1
0000,
                                                                 "CI_per_hundred_thousand" = NVISITS/NMEM
B *100000)

flu_by_month_regions_only |> ggplot(aes(x = MONTH, y = CI_per_thousand, group = part)) +
  geom_line(aes(color = part)) + scale_y_continuous(trans = "log10") +
  facet_wrap(~YEAR) + scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG",
"SEPT",

"OCT", "NOV", "DEC")) + theme_bw() +theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
ggtitle("Flu Cases Over Time \n All Ages") +ylab("Cases per 1000 People") +  labs(color = "Region")

#need to figure out where data is missing

flu_vax <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/flu_data/Influenza_Vac
cination_Coverage_for_All_Ages__6__Months_.csv")

flu_vax_filt <- flu_vax |>
  filter(`Season/Survey Year` %in% c("2009-10", "2010-11", "2011-12",
                                                                   "2012-13", "2013-14", "2014-15",
                                                                   "2015-16", "2016-17", "2017-18",
                                                                   "2018-19")) |>
  filter(Dimension %in% c("≥6 Months","6 Months - 17 Years", "≥18 Years" , "6 Months - 4 Years" ,
                        "5-12 Years",  "13-17 Years")) |>
  mutate(state = tolower(Geography)) |>
  filter(state %in% state_parts_2$state) |>
  select(-`Geography Type`, -Vaccine, -FIPS, -`Dimension Type`)

#need to put the years on the same year scale as the cases data
flu_vax_filt <- flu_vax_filt |> mutate(year_lower = substr(`Season/Survey Year`, 1,4),
                       year_higher = paste0("20",str_sub(`Season/Survey Year`, -2, -1)),
                       year = ifelse(Month %in% c(7,8,9,10,11,12), year_lower,
                              ifelse(Month %in% c(1,2,3,4,5), year_higher, NA)),
                       state = tolower(Geography),
                       coverage = as.numeric(`Estimate (%)`))

#add region to this
flu_vax_filt <- left_join(flu_vax_filt, state_parts_2, by= "state")
flu_vax_filt_2 <- flu_vax_filt |> mutate(STATE = Geography, YEAR = as.numeric(year)) |>
  select(Month, STATE, Dimension, `Estimate (%)`,YEAR, part) |> filter(YEAR != 2009)

#for the sake of the geographic analysis, I will show flu vaccination rates in 5-9 year olds by november
of each year in each state
fluvax_map <- flu_vax_filt_2 |>
  filter(Dimension == "5-12 Years", Month == 11 ) |>
  mutate(region = tolower(STATE)) |>
  left_join(us_df)

us_base + geom_polygon(data = fluvax_map, aes(fill = as.numeric(`Estimate (%)`)))+
  geom_polygon(color = "gray", fill = NA) + theme(
  axis.text = element_blank(),
  axis.line = element_blank(),
  axis.ticks = element_blank(),
  panel.border = element_blank(),
  panel.grid = element_blank(),
  axis.title = element_blank()
  ) + facet_wrap(~YEAR) + scale_fill_gradient(low = "#FFFFFF", high = "#F05E16" ) + ggtitle("November Flu
Vaccination Coverage") + labs(fill = "Estimated Flu Vaccination %")
```

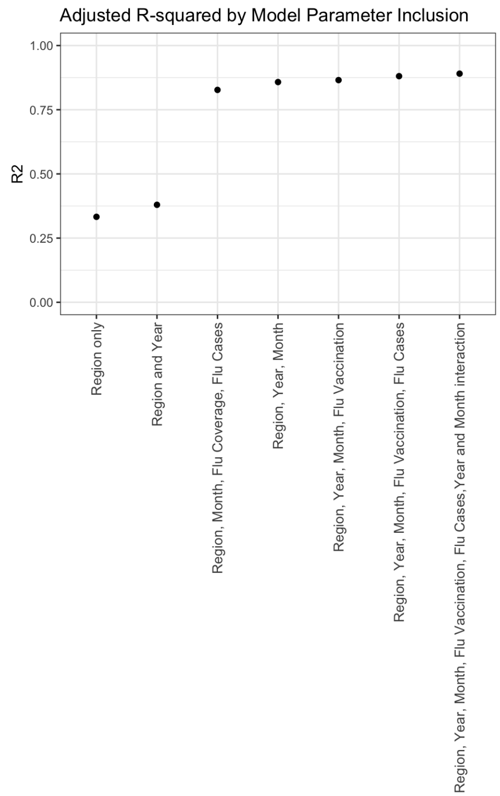
```
#now take >6 months, which is vaccination in all ages. Do 5-12 separately
flu_vax_filt_allages <- flu_vax_filt_2 |> filter(Dimension == "≥6 Months")

#get the cohort membership by state from the original dataframe
coh_fluvax_allages <- aggregate(NMEMB ~ STATE + YEAR, data = coh, sum)
flu_vax_rescale_allages <- left_join(flu_vax_filt_allages, coh_fluvax_allages) |>
  mutate(sample_vaxxed = as.numeric(`Estimate (%)`)*1/100*NMEMB)
#aggregate the members and sample vaxxed
flu_vax_rescale_allages_memb <- aggregate(NMEMB ~ part + YEAR + Month, data = flu_vax_rescale_allages, su
m)
flu_vax_rescale_allages_vax <- aggregate(sample_vaxxed ~ part + YEAR + Month, data = flu_vax_rescale_alla
ges, sum)
flu_vax_rescale_allages_comb <- left_join(flu_vax_rescale_allages_vax, flu_vax_rescale_allages_memb) |> m
utate(flu_coverage = sample_vaxxed / NMEMB* 100)


#aggregate(NMEMB ~ STATE + YEAR, data = by_month_regions, sum)
#aggregate(NMEMB ~ STATE + YEAR, data = flu_by_month_regions, sum)
#these are different by a factor of 12, which makes sense because in one case I am summing over all the m
onths.
#I think this is still what I want, but need to check with stephen. Shouldn't affect scaling though cuz i
ts just off by a factor of 12 which will come out regardless

#do the same thing for the 5-9 age group
flu_vax_filt_512 <- flu_vax_filt_2 |> filter(Dimension == "5-12 Years")
coh_fluvax_59 <- aggregate(NMEMB~ STATE + YEAR + AGEGRP, data = coh, sum) |> filter(AGEGRP == "05_09")
flu_vax_rescale_5 <- left_join(flu_vax_filt_512, coh_fluvax_59) |> mutate(sample_vaxxed = as.numeric(`Est
imate (%)`)*1/100*NMEMB)
flu_vax_rescale_5_memb <- aggregate(NMEMB ~ part + YEAR + Month, data = flu_vax_rescale_5, sum)
flu_vax_rescale_5_vax <- aggregate(sample_vaxxed ~ part + YEAR + Month, data = flu_vax_rescale_5, sum)
flu_vax_rescale_5_comb <- left_join(flu_vax_rescale_5_memb, flu_vax_rescale_5_vax) |>
  mutate(flu_coverage = sample_vaxxed / NMEMB * 100)

#plot the strep and flu vaccine data ontop of one another for all ages, and for 5-9 age group specificall
y
flu_vax_rescale_allages_comb <- flu_vax_rescale_allages_comb |> mutate(MONTH = Month) |> select(-Month)
joint2 <- by_month_regions_only |> select(-NVISITS, -NMEMB) |> left_join(flu_vax_rescale_allages_comb) |>
  select(-sample_vaxxed, -NMEMB) |> mutate(Strep_CI_per_ten_thousand = CI_per_thousand*10)
joint2_pivot <- joint2|> pivot_longer(cols = c("Strep_CI_per_ten_thousand", "flu_coverage"))
fluvax_strep_allages_plot <- joint2_pivot |> ggplot(aes(x= MONTH, y = value, color = part)) +
  geom_line(aes(linetype = name)) +
  facet_wrap(~YEAR) + scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG",
"SEPT",

"OCT", "NOV", "DEC")) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + ggtitle("Strep Pharyngitis Rates
and Flu Vaccination Over Time  \n All Ages, scaled to MarketScan maybe") + labs(linetype = "Value Plotte
d") + labs(color = "Region") + scale_linetype_discrete(labels = c("Flu Vaccination %", "Strep cases per 1
0,000"))


#for 5-9 age group
by_month_regions_5 <- by_month_regions |> filter(AGEGRP %in% c("05_09"))
by_month_regions_5_vis <- aggregate(NVISITS ~ part  + YEAR + MONTH,
                                       dat = by_month_regions_5,
                                       sum)
by_month_regions_5_memb <- aggregate(NMEMB ~ part  + YEAR + MONTH,
                                       dat = by_month_regions_5,
                                       sum)
by_month_regions_5 <- left_join(by_month_regions_5_vis, by_month_regions_5_memb)
by_month_regions_5 <- by_month_regions_5 |> mutate("CI_per_thousand" = NVISITS/NMEMB *1000)
#fluvax data
flu_vax_rescale_5_comb <- flu_vax_rescale_5_comb |> mutate(MONTH = Month) |> select(-Month)
joint3 <- by_month_regions_5 |> select(-NVISITS, -NMEMB) |> left_join(flu_vax_rescale_5_comb) |>
  select(-sample_vaxxed, -NMEMB) |> mutate(Strep_CI_per_ten_thousand = CI_per_thousand*10, Strep_CI_per_f
ive_thousand = CI_per_thousand*5, Strep_CI_per_two_thousand = CI_per_thousand*2)
joint3_pivot <- joint3 |> pivot_longer(cols = c("Strep_CI_per_two_thousand", "flu_coverage"))
fluvax_strep_59_plot <- joint3_pivot |> ggplot(aes(x= MONTH, y = value, color = part)) +
```

```
  geom_line(aes(linetype = name)) +
  facet_wrap(~YEAR) + scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG",
"SEPT",

"OCT", "NOV", "DEC")) +theme_bw() + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
ggtitle("Strep Pharyngitis Rates and Flu Vaccination Over Time  \n 5-9 Age Group") + labs(linetype = "Val
ue Plotted") + labs(color = "Region") + scale_linetype_discrete(labels = c("Flu Vaccination %", "Strep ca
ses per 2,000"))

fluvax_strep_allages_plot
fluvax_strep_59_plot

childvax_data <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/vax_data/Vaccina
tion_Coverage_among_Young_Children__0-35_Months_.csv")
pcv_vax <- childvax_data |> filter(Vaccine == "PCV")
pcv_4d_35m <- pcv_vax |> filter(Dose == "≥4 Doses") |> filter(Dimension == "35 Months")
pcv_4d_35m <- pcv_4d_35m |> filter(`Birth Year/Birth Cohort` %in% c('2011', '2012', '2013',
                                                    '2014', '2015', '2016', '2017', '2018')) |>
  mutate(year = as.numeric(`Birth Year/Birth Cohort`), region = tolower(Geography))
pcv_map <- pcv_4d_35m |> filter(year %in% c(2011,2012,2013)) |> left_join(us_df)
pcv_over_time_4d35m <- us_base + geom_polygon(data = pcv_map, aes(fill = `Estimate (%)`))+
  geom_polygon(color = "gray", fill = NA) + theme(
  axis.text = element_blank(),
  axis.line = element_blank(),
  axis.ticks = element_blank(),
  panel.border = element_blank(),
  panel.grid = element_blank(),
  axis.title = element_blank()
  ) + facet_wrap(~year) + scale_fill_gradient(low = "#FFFFFF", high = "#800000" ) + ggtitle("PCV Vaccinat
ion by 35 months by birth cohort")
pcv_over_time_4d35m


pcv_grouping <- pcv_4d_35m |> filter(`Birth Year/Birth Cohort` %in% c("2011", "2012", "2013")) |>
  mutate(STATE = Geography, YEAR = as.numeric(`Birth Year/Birth Cohort`)) |>
  select(Vaccine, STATE, `Estimate (%)`, YEAR, `Sample Size`)
pcv_memb_ag <- coh |> filter(YEAR %in% c("2011", "2012", "2013"), AGEGRP == "00_04")
pcv_memb_ag <- aggregate(NMEMB ~ STATE + AGEGRP + YEAR, dat = pcv_memb_ag, sum)
pcv_memb_ag <- pcv_memb_ag |> mutate(YEAR = as.numeric(YEAR))
pcv_grouped <- left_join(pcv_grouping, pcv_memb_ag)
pcv_grouped <- pcv_grouped |> mutate(sample_vaxxed = `Estimate (%)`*0.01 *NMEMB)
pcv_grouped <- pcv_grouped |> mutate(state = tolower(STATE))
pcv_grouped <- left_join(pcv_grouped, state_parts_2)
pcv_grouped_2016 <- pcv_grouped |> filter(YEAR == 2011) |> na.omit() #NAs are for regions that aren't sta
tes; ok to omit
pcv_grouped_2017 <- pcv_grouped |> filter(YEAR %in% c(2011, 2012)) |> na.omit()
pcv_grouped_2018 <- pcv_grouped |> na.omit()

pcv_grouped_2016_vaxxed <- aggregate(sample_vaxxed ~ part, data = pcv_grouped_2016, sum)
pcv_grouped_2016_memb <- aggregate(NMEMB ~ part, data = pcv_grouped_2016, sum)
pcv_2016_ag_region <- left_join(pcv_grouped_2016_memb, pcv_grouped_2016_vaxxed) |> mutate(pcv_coverage =
sample_vaxxed / NMEMB * 100, YEAR = 2016)

pcv_grouped_2017_vaxxed <- aggregate(sample_vaxxed ~ part, data = pcv_grouped_2017, sum)
pcv_grouped_2017_memb <- aggregate(NMEMB ~part, data = pcv_grouped_2017, sum)
pcv_2017_ag_region <- left_join(pcv_grouped_2017_memb, pcv_grouped_2017_vaxxed) |> mutate(pcv_coverage =
sample_vaxxed / NMEMB * 100, YEAR = 2017)

pcv_grouped_2018_vaxxed <- aggregate(sample_vaxxed ~ part, data = pcv_grouped_2018, sum)
pcv_grouped_2018_memb <- aggregate(NMEMB ~part, data = pcv_grouped_2018, sum)
pcv_2018_ag_region <- left_join(pcv_grouped_2018_memb, pcv_grouped_2018_vaxxed) |> mutate(pcv_coverage =
sample_vaxxed / NMEMB * 100, YEAR = 2018)

all_pcv <- rbind(pcv_2016_ag_region, pcv_2017_ag_region, pcv_2018_ag_region) |> select(part, pcv_coverag
e, YEAR) #coverage for 5-9 year olds

#plot it ontop of strep cases
#look by region over time by age group
by_year_age_region_visits <- aggregate(NVISITS ~ part + AGEGRP + PRIMARYCOND + YEAR, dat = by_year_age, s
um)
by_year_age_region_members <- aggregate(NMEMB ~ part + AGEGRP + PRIMARYCOND + YEAR, dat = by_year_age, su
m)
```

```
by_year_age_region <- left_join(by_year_age_region_visits,by_year_age_region_members)
by_year_age_region <- by_year_age_region |> mutate(CI_per_hundred = NVISITS/NMEMB*100)
strep_for_pcv <- by_year_age_region |> filter(AGEGRP == "05_09", YEAR %in% c(2016, 2017, 2018)) |>
  mutate(Strep_CI_per_hundred = CI_per_hundred) |> select(part, YEAR, Strep_CI_per_hundred)
strep_pcv <- left_join(strep_for_pcv, all_pcv) |> mutate(Strep_CI_per_five_hundred = Strep_CI_per_hundred
*5) |>
  select(-Strep_CI_per_hundred)
strep_pcv_pivoted <- pivot_longer(strep_pcv, cols = c("Strep_CI_per_five_hundred", "pcv_coverage"))
# strep_pcv_pivoted|>
#   ggplot(aes(x= YEAR, y = value, color = part)) +
#   geom_line(aes(linetype = name)) + scale_x_discrete(limits = c(2016, 2017, 2018))

#maybe separate by region, do barplots
midwest <- strep_pcv_pivoted |> filter(part == "midwest") |> ggplot(aes(x = YEAR, y = value, fill = nam
e)) + geom_bar(stat = "identity", color = "black", position = position_dodge()) + labs(fill = "Value") +
scale_fill_manual(values = c("#800000", "#C21807" ), labels = c("PCV Vaccination %", "Strep Cases per 50
0")) + ylim(0,100) + ggtitle("PCV Coverage and Strep Pharyngitis Cases Midwest") + theme_bw()

northeast <- strep_pcv_pivoted |> filter(part == "northeast") |> ggplot(aes(x = YEAR, y = value, fill = n
ame)) + geom_bar(stat = "identity", color = "black", position = position_dodge()) + labs(fill = "Value")
+ scale_fill_manual(values = c("#800000", "#658354" ), labels = c("PCV Vaccination %", "Strep Cases per 5
00")) + ylim(0,100) + ggtitle("PCV Coverage and Strep Pharyngitis Cases Northeast") + theme_bw()

west <- strep_pcv_pivoted |> filter(part == "west") |> ggplot(aes(x = YEAR, y = value, fill = name)) + ge
om_bar(stat = "identity", color = "black", position = position_dodge()) + labs(fill = "Value") + scale_fi
ll_manual(values = c("#800000", "#7852A9" ), labels = c("PCV Vaccination %", "Strep Cases per 500")) + yl
im(0,100) + ggtitle("PCV Coverage and Strep Pharyngitis Cases West") + theme_bw()

south <- strep_pcv_pivoted |> filter(part == "south") |> ggplot(aes(x = YEAR, y = value, fill = name)) +
geom_bar(stat = "identity", color = "black", position = position_dodge()) + labs(fill = "Value") + scale_
fill_manual(values = c("#800000", "#6699CC" ), labels = c("PCV Vaccination %", "Strep Cases per 500")) +
ylim(0,100) + ggtitle("PCV Coverage and Strep Pharyngitis Cases South") + theme_bw()

grid.arrange(midwest, northeast, west, south, ncol = 2, nrow =2)
```

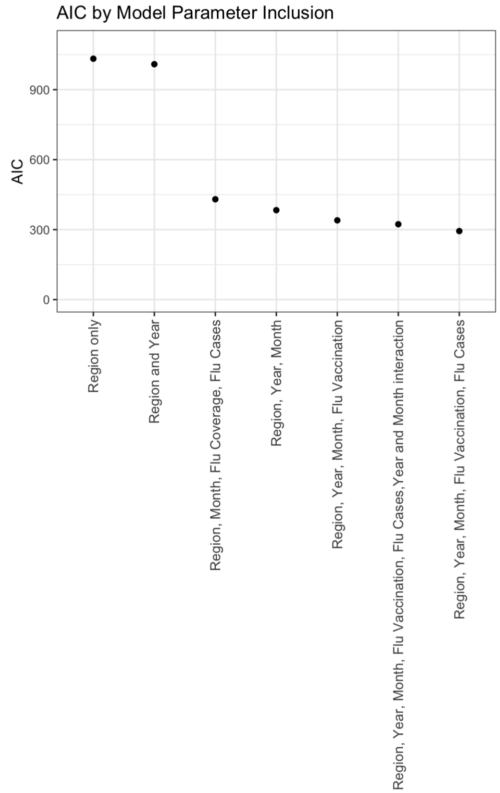
Models

I will fit a model to better understand drivers of regional patterns of streptococcal pharyngitis. I will first consider all age groups across all years of data (2010-2018), and with possible covariates region, year, month, flu vaccination (monthly), and flu cases. I am using a linear regression model because streptococcal pharyngitis cases are already reported as total counts for a given year and month, so with large enough amounts of data, the errors will follow a normal distribution. In order to select an optimal model, I tested multiple combinations of parameters, and compared the R^2 values of these models to start. I used an adjusted R^2, which is calculated using a penalty for parameters $R^2_{adj} = R^2 - \frac{p(1-R^2)}{(n-p-1)}$ where $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})}{\sum_{i=1}^{n}(y_i - \bar{y})}$.

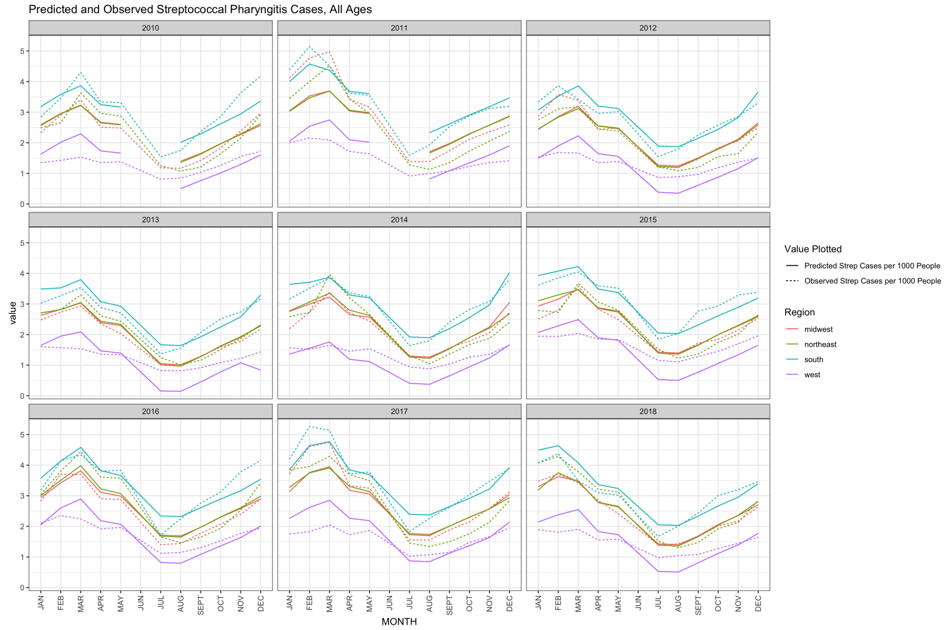


We can see that R^2_{adj} largely increases as number of parameters increase, although not always (the model with Region, Year, and Month has a higherR^2_{adj} than the model with Region, Month, Flu coverage, and Flu cases). Larger values of R^2_{adj} mean that the model accounts for a greater proportion of the variability of the outcome. The model with the highest R^2_{adj} contains region, year, month, flu vaccination, flu cases, and a statistical interaction term bewteen year and month, and has an R^2_{adj} of 0.89, The model with the second highest R^2_{adj} contains region, year, month, flu vaccination, and flu cases, without a statistical interaction term between year and month. This model has an R^2_{adj} of 0.88.

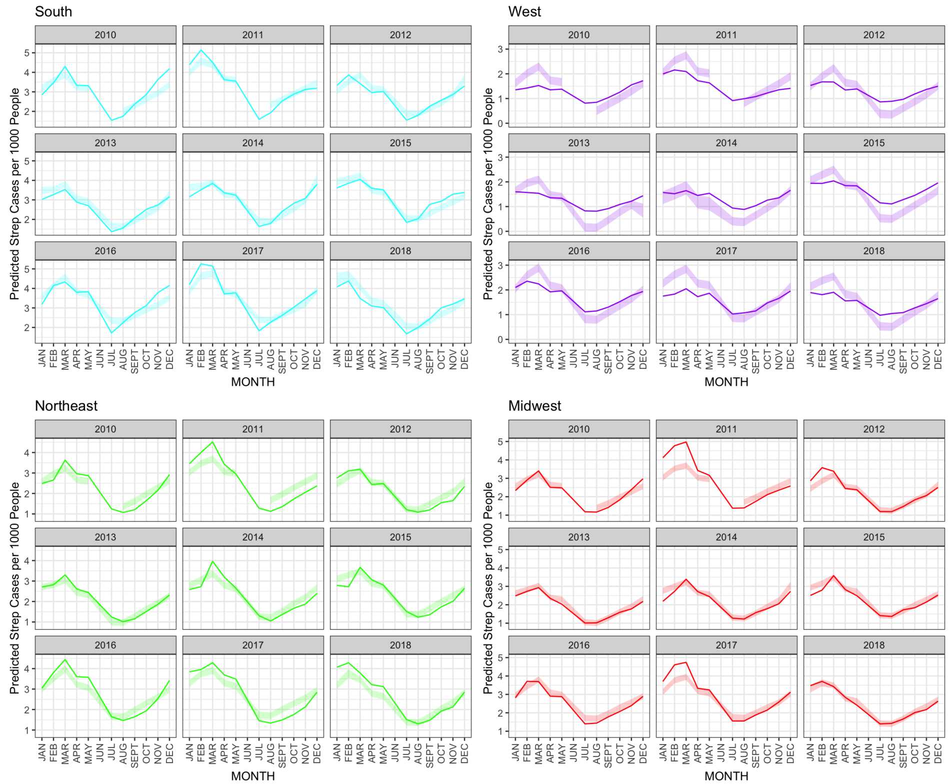
We can also use the Akaike Information Criterion (AIC), calculated as $AIC = -2log(\hat{L}) + 2(p+1)$ where $\hat{L}$ is the maximum likelihood of the model, and p is number of parameters. The rationale for this criterion is that the likelihood must go up by more than is justified by the additional parameters added. Lower AIC values are better.



The model with the lowest AIC contains region, year, month, flu vaccination, and flu cases, without a statistical interaction between year and month. This model also has the second highest R^2_{adj} so, it is likely the best fit of the available models. We can visualize our model's predictions against the observed data.

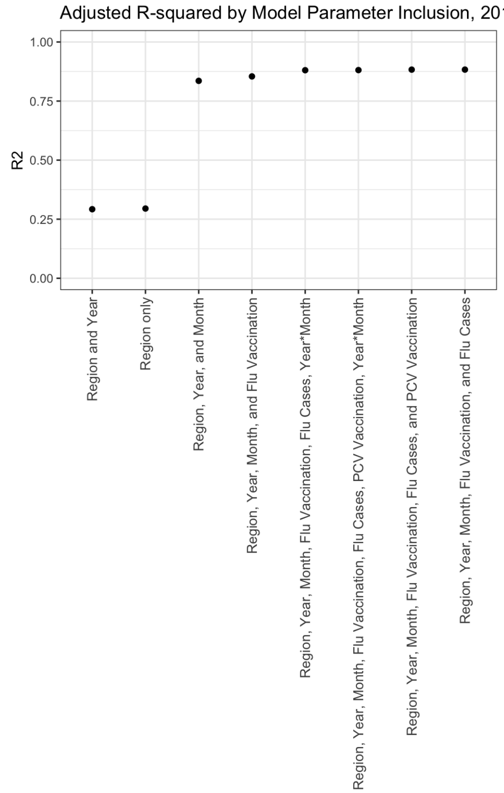


The model (solid line) generally tracks well with observed cases of streptococcal pharyngitis. We can next plot the model predictions with corresponding 95% confidence intervals against observed cases for each region separately.

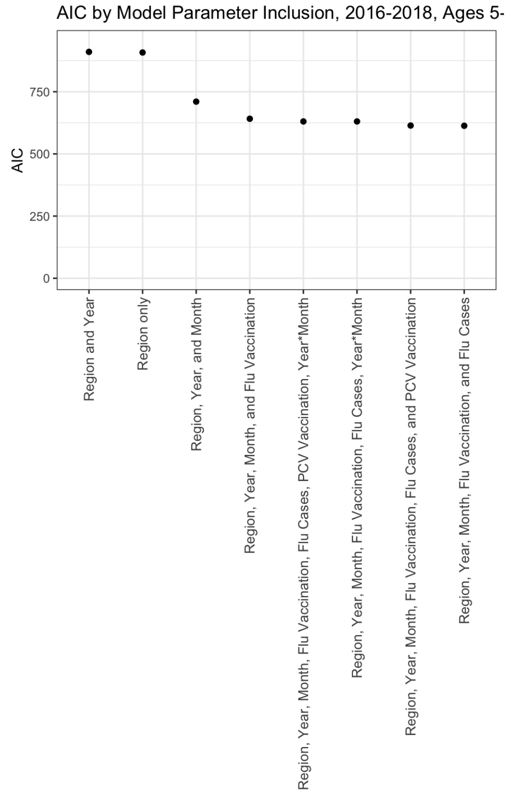


At this finer resolution, the model appears to predict cases in the South well, though it does not always capture the steep rise in cases in the fall and early winter. It somewhat overpredicts seasonal changes in the West (it predicts more cases in the beginning of the year and fewer in the early winter than are observed), and it predicts the Northeast and Midwest relatively faithfully.

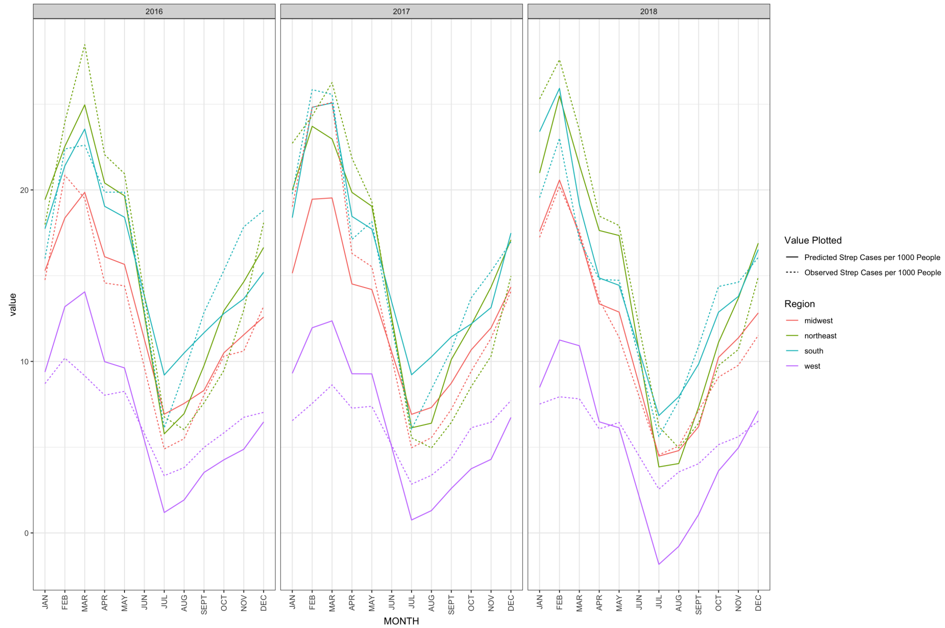
We next apply a similar strategy also including PCV vaccination data. Because PCV vaccination from the CDC is broken down by birth cohort, we include data from the 2011, 2012, and 2013 birth cohorts and compare children in these cohorts who received all 4 doses to cases of streptococcal pharyngitis and flu in children ages 5-9 in years 2016, 2017, and 2018, when the children from these birth cohorts enter this age range. Flu vaccination data are taken from chidlren in the 5-12 year old age range. As before, we first compare R^2_{adj} for linear models with different parameters.



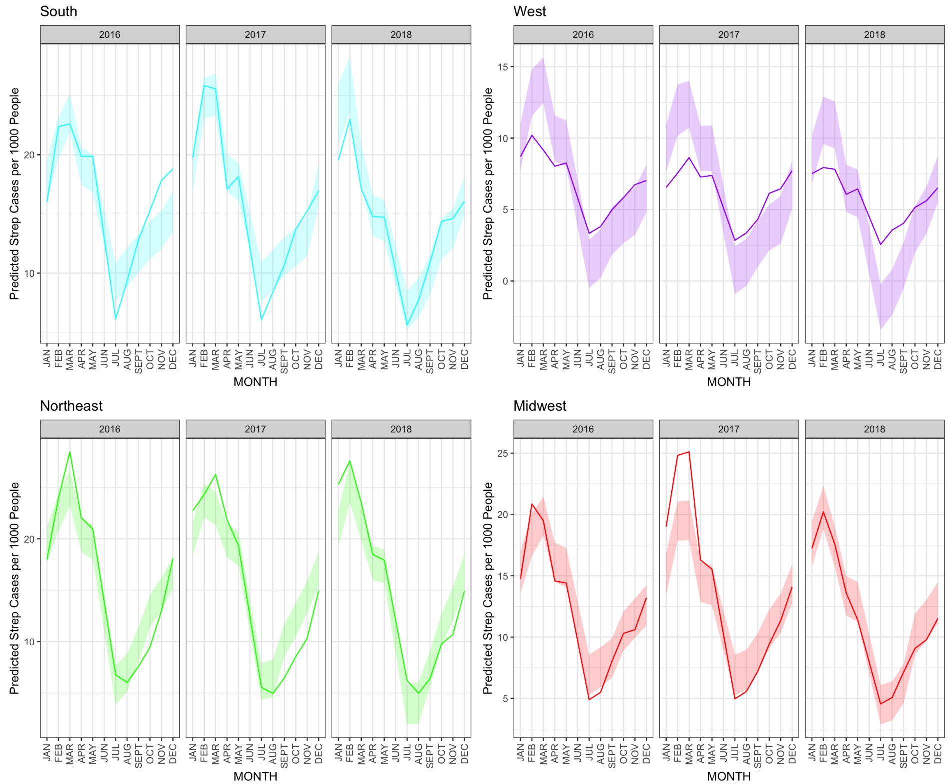
The model with the highest value of R^2_{adj} here is not the model with the most parameters, but rather includes region, year, month, flu vaccination and flu cases. The R^2_{adj} for this model is 0.88, indicating that it explains about 88% of the variability in streptococcal pharyngitis cases. Notably, models including PCV vaccination have lower R^2_{adj} values, indicating that this covariate does not help to explain variation in cases of streptococcal pharyngitis. We next compare AIC for each model.



Here, the lowest AIC also corresponds to the same model with the highest R^2_{adj}, indicating that this is the best available model. We can again visualize our model's predictions against observed data.



The models once again track reasonably well with observed data. We can zoom in on predictions in each region as before.



Here we see similar trends as before, which makes sense given that we have essentially fit the same model just in years 2016-2018, and with only 5-9 year olds (the age group with the highest burden of disease).

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(stringr)
library(ggplot2)
library(usmap)
library(gridExtra)
library(maps)
library(mapdata)
library(ggmap)
library(lmtest)

#Read in data on cases
dat <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/StrepPharyngitis/output/Ge
oVisits.csv")

#Read in data on membership
coh <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/StrepPharyngitis/output/Ge
oCohort.csv")
#this dataframe has population by sex, age group, and state but not by month. So need to add it to the ot
her dataframe once already collapsed by year.

#get data by year, rather than by month:
by_year <- aggregate(NVISITS ~ YEAR + STATE + SEX + AGEGRP + PRIMARYCOND, dat, sum)
#we can then join this with the population data from cohort
by_year <- left_join(by_year, coh)

# by_year |> group_by(YEAR) |> summarize(total_vis = sum(NVISITS), total_memb = sum(NMEMB), CI = round(to
tal_vis/total_memb,4)) |> ggplot(aes(x=YEAR, y = CI)) + geom_bar(stat = "identity", fill = "steelblue") +
theme_bw() + geom_text(aes(label=CI), vjust=1.6, color="white", size=2.0) + ylab("Cases per person") +  g
gtitle("Total Strep Pharyngitis Visits per Member by Year US")


#let's summarize cumulative incidence by state across years to start
by_state <- by_year |>
  group_by(YEAR, STATE) |>
  summarize(visits = sum(NVISITS), members = sum(NMEMB), CI = visits/members)

#we now manipulate the data slightly to show CI per hundred, and make the state names match the mapping d
ataframe
strep_all <- by_state |> mutate(region = tolower(STATE), CI_per_hundred = CI*100) |>
  select(region, CI_per_hundred)



#now we add region designations for "northeast", "south", "midwest", and "west"
#add regions to this dataframe
#make a function that converts lists of state abbreviations to lists of state names
to_statename <- function(list){
  new_list <- c()
  for(i in 1:length(list)){
    name <- state.name[grep(list[i], state.abb)]
    new_list <- append(new_list, name)
  }
  new_list
}

northeast_states <- tolower(to_statename(.northeast_region))
midwest_states <- tolower(to_statename(.midwest_region))
south_states <- tolower(to_statename(.south_region))
west_states <- tolower(to_statename(.west_region))

northeast_df <- data.frame(region = northeast_states, part = "northeast")
midwest_df <- data.frame(region = midwest_states, part = "midwest")
south_df <- data.frame(region = south_states, part = "south")
west_df <- data.frame(region = west_states, part = "west")
#will put dc in the south because maryland and viriginia are
```

```
dc_df <- data.frame(region = "washington dc", part = "south")


state_parts <- rbind(northeast_df, midwest_df, south_df, west_df, dc_df)


strep_all_region <- left_join(strep_all, state_parts, by = "region")

#now we look at the data by age group across all regions
by_year <- by_year |> mutate("state" = STATE)
by_year_age_visits <- aggregate(NVISITS ~ YEAR + state + AGEGRP + PRIMARYCOND, dat = by_year, sum)
by_year_age_membs <- aggregate(NMEMB ~ YEAR + state + AGEGRP + PRIMARYCOND, dat = by_year, sum)
by_year_age <- left_join(by_year_age_visits, by_year_age_membs)
by_year_age <- by_year_age |> mutate(CI_per_hundred = NVISITS/NMEMB *100, state = tolower(state))

#add in region just in case?
state_parts_2 <- state_parts
names(state_parts_2)[1] <- "state"

by_year_age <- left_join(by_year_age, state_parts_2)
#make a plot of trends over time by state


country_by_age <- left_join(aggregate(NVISITS ~ AGEGRP + YEAR, dat = by_year_age, sum), aggregate(NMEMB ~
AGEGRP + YEAR, dat = by_year_age, sum))
country_by_age <- country_by_age |> mutate(CI_per_hundred = NVISITS/NMEMB * 100)

# country_by_age |> ggplot(aes(YEAR, CI_per_hundred, group = AGEGRP)) + geom_line(aes(color = AGEGRP)) +
ggtitle("Streptococcal Pharyngitis in the US from 2010-2018 \n By Age Group") + theme_bw() + ylab("Cases
per 100 People") + labs(color = "Age Group")



states_indiv_region <- strep_all_region |> group_by(YEAR, region) |> ggplot(aes(YEAR, CI_per_hundred, gro
up=region)) + geom_line(aes(col = part)) + ggtitle("Streptococcal Pharyngitis In All States by Region") +
ylab("Cases per 100 People") + theme_bw() + labs(color = "Region")

#now group them by region and just report 1 value per region

by_state_2 <- by_state |> mutate(state = tolower(STATE))
strep_region_visits <- left_join(by_state_2, state_parts_2, by = "state")
strep_region_visits_agg <- aggregate(visits~ part + YEAR, dat = strep_region_visits, sum)
strep_region_members_agg <- aggregate(members~ part + YEAR, dat = strep_region_visits, sum)
strep_region_joined <- left_join(strep_region_visits_agg, strep_region_members_agg)
strep_region_joined <- strep_region_joined |> mutate(CI_per_hundred = visits/members * 100)

per_region<- strep_region_joined |> group_by(YEAR, part) |> ggplot(aes(YEAR, CI_per_hundred, group=part))
+ geom_line(aes(col = part)) + ggtitle("Streptococcal Pharyngitis By Region") +ylab("Cases per 100 Peopl
e") + theme_bw() + labs(color = "Region")

#grid.arrange(states_indiv_region, per_region, ncol=2)

#try by month; first just plot data by region for all age groups together
#should be able to just merge these two because the members are stable over the year
by_month_all <- left_join(dat, coh)

#consolidate across sex and check if this is what you needed to do earlier on as well;
#to consolidate across sex, will just need to add so that should be fine
by_month_age_vis <- aggregate(NVISITS ~ MONTH + STATE + AGEGRP + YEAR, dat = by_month_all, sum)
memb_no_sex <- aggregate(NMEMB ~  STATE +AGEGRP + YEAR, dat = coh, sum)
by_month_no_sex <- left_join(by_month_age_vis, memb_no_sex)
by_month_no_sex <- by_month_no_sex |> mutate("CI_per_thousand" = NVISITS/NMEMB*1000)



by_month_no_sex_lowercase <- by_month_no_sex |> mutate("state" = tolower(STATE))
by_month_regions <- left_join(by_month_no_sex_lowercase, state_parts_2, by = "state")
by_month_regions_only_vis <- aggregate(NVISITS ~ part  + YEAR + MONTH,
                                       dat = by_month_regions,
                                       sum)
by_month_regions_only_memb <- aggregate(NMEMB ~ part  + YEAR + MONTH,
                                        dat = by_month_regions,
```

```
                                    sum)


by_month_regions_only <- left_join(by_month_regions_only_vis, by_month_regions_only_memb)
by_month_regions_only <- by_month_regions_only |> mutate("CI_per_thousand" = NVISITS/NMEMB *1000)
# by_month_regions_only |> ggplot(aes(x = MONTH, y = CI_per_thousand, group = part)) +
#   geom_line(aes(color = part)) +
#   facet_wrap(~YEAR) + ggtitle("Steptococcal Pharyngitis Cases Incidence by Month \n All Ages, by Regio
n") + labs(color = "Region") + ylab("Cases per 1,000 People") +scale_x_discrete(limits = c("JAN", "FEB",
"MAR", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT",
#
"OCT", "NOV", "DEC")) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))



#we now load in US mapping data from the maps package
us_df <- map_data("state")

#this creates a base US map to build upon
us_base <- ggplot(data = us_df, mapping = aes(x = long, y = lat, group =group))+
  coord_fixed(1.3) + geom_polygon(color="black", fill = "gray")

#we join our dataframe with strep info with our mapping dataframe
us_strep_all <- inner_join(us_df, strep_all, by = "region")
# us_strep_rates_over_time <- us_base + geom_polygon(data = us_strep_all, aes(fill = CI_per_hundred))+
#   geom_polygon(color = "gray", fill = NA) + theme(
#   axis.text = element_blank(),
#   axis.line = element_blank(),
#   axis.ticks = element_blank(),
#   panel.border = element_blank(),
#   panel.grid = element_blank(),
#   axis.title = element_blank()
#   ) + facet_wrap(~YEAR) + scale_fill_gradient(low = "#FFFFFF", high = "#FF0000" ) + ggtitle("Cumulative
Incidence of Streptococcal Pharyngitis from 2010-2018 \n All Ages") + labs(fill = "Cases per 100 People")
# us_strep_rates_over_time


#making maps for age group trends over time
by_year_age_formap <- by_year_age
names(by_year_age_formap)[2] = "region"
us_strep_byage_map <- inner_join(us_df, by_year_age_formap, by = "region")
#attempt facet_grid with year as the column and age group as the row
# us_strep_byage_map_gg <- us_base +
#   geom_polygon(data = us_strep_byage_map, aes(fill = CI_per_hundred)) +
#     geom_polygon(color = "gray", fill = NA) +
#   theme(
#   axis.text = element_blank(),
#   axis.line = element_blank(),
#   axis.ticks = element_blank(),
#   panel.border = element_blank(),
#   panel.grid = element_blank(),
#   axis.title = element_blank()
#   ) + scale_fill_gradient(low = "#FFFFFF", high = "#FF0000" ) + ggtitle("Cumulative Incidence of Strept
ococcal Pharyngitis from 2010-2018 \n By Age Group") + labs(fill = "Cases per 100 People")
#
# us_strep_byage_map_gg + facet_grid(rows = vars(AGEGRP), cols = vars(YEAR) )
#now for flu case data
flu_dat <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/StrepPharyngitis/outpu
t/GeoVisitsFlu.CSV")
flu_coh <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/StrepPharyngitis/outpu
t/GeoCohortFlu.csv") #this is the same as strep data

#do same thing that did for strep cases to start
#should be able to just merge these two because the members are stable over the year
by_month_all_flu <- left_join(flu_dat, flu_coh)

flu_by_month_age_vis <- aggregate(NVISITS ~ MONTH + STATE + AGEGRP + YEAR, dat = by_month_all_flu, sum)
flu_memb_no_sex <- aggregate(NMEMB ~  STATE +AGEGRP + YEAR, dat = flu_coh, sum)
flu_by_month_no_sex <- left_join(flu_by_month_age_vis, flu_memb_no_sex)
flu_by_month_no_sex <- flu_by_month_no_sex |>
  mutate("CI_per_thousand" = NVISITS/NMEMB*1000) |>
  mutate("CI_per_ten_thousand" =NVISITS/NMEMB*10000 )
```

```
flu_by_month_no_sex_lowercase <- flu_by_month_no_sex |> mutate("state" = tolower(STATE))
flu_by_month_regions <- left_join(flu_by_month_no_sex_lowercase, state_parts_2, by = "state")
flu_by_month_regions_only_vis <- aggregate(NVISITS ~ part  + YEAR + MONTH,
                                       dat = flu_by_month_regions,
                                       sum)
flu_by_month_regions_only_memb <- aggregate(NMEMB ~ part  + YEAR + MONTH,
                                       dat = flu_by_month_regions,
                                       sum)

flu_by_month_regions_only <- left_join(flu_by_month_regions_only_vis, flu_by_month_regions_only_memb)
flu_by_month_regions_only <- flu_by_month_regions_only |> mutate("CI_per_thousand" = NVISITS/NMEMB *1000,
                                                                 "CI_per_ten_thousand" = NVISITS/NMEMB *1
0000,
                                                                 "CI_per_hundred_thousand" = NVISITS/NMEM
B *100000)

# flu_by_month_regions_only |> ggplot(aes(x = MONTH, y = CI_per_thousand, group = part)) +
#   geom_line(aes(color = part)) + scale_y_continuous(trans = "log10") +
#   facet_wrap(~YEAR) + scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AU
G", "SEPT",
#
"OCT", "NOV", "DEC")) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + ggtitle("Fl
u Cases Over Time \n All Ages")

#need to figure out where data is missing

flu_vax <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/flu_data/Influenza_Vac
cination_Coverage_for_All_Ages__6__Months_.csv")

flu_vax_filt <- flu_vax |>
  filter(`Season/Survey Year` %in% c("2009-10", "2010-11", "2011-12",
                                                                   "2012-13", "2013-14", "2014-15",
                                                                   "2015-16", "2016-17", "2017-18",
                                                                   "2018-19")) |>
  filter(Dimension %in% c("≥6 Months","6 Months - 17 Years", "≥18 Years" , "6 Months - 4 Years" ,
                          "5-12 Years",  "13-17 Years")) |>
  mutate(state = tolower(Geography)) |>
  filter(state %in% state_parts_2$state) |>
  select(-`Geography Type`, -Vaccine, -FIPS, -`Dimension Type`)

#need to put the years on the same year scale as the cases data
flu_vax_filt <- flu_vax_filt |> mutate(year_lower = substr(`Season/Survey Year`, 1,4),
                       year_higher = paste0("20",str_sub(`Season/Survey Year`, -2, -1)),
                       year = ifelse(Month %in% c(7,8,9,10,11,12), year_lower,
                                     ifelse(Month %in% c(1,2,3,4,5), year_higher, NA)),
                       state = tolower(Geography),
                       coverage = as.numeric(`Estimate (%)`))

#add region to this
flu_vax_filt <- left_join(flu_vax_filt, state_parts_2, by= "state")
flu_vax_filt_2 <- flu_vax_filt |> mutate(STATE = Geography, YEAR = as.numeric(year)) |>
  select(Month, STATE, Dimension, `Estimate (%)`,YEAR, part) |> filter(YEAR != 2009)

#for the sake of the geographic analysis, I will show flu vaccination rates in 5-9 year olds by november
of each year in each state
fluvax_map <- flu_vax_filt_2 |>
  filter(Dimension == "5-12 Years", Month == 11 ) |>
  mutate(region = tolower(STATE)) |>
  left_join(us_df)

# us_base + geom_polygon(data = fluvax_map, aes(fill = as.numeric(`Estimate (%)`)))+
#   geom_polygon(color = "gray", fill = NA) + theme(
#   axis.text = element_blank(),
#   axis.line = element_blank(),
#   axis.ticks = element_blank(),
#   panel.border = element_blank(),
#   panel.grid = element_blank(),
#   axis.title = element_blank()
#   ) + facet_wrap(~YEAR) + scale_fill_gradient(low = "#FFFFFF", high = "#F05E16" ) + ggtitle("November F
lu Vaccination Coverage")
#
```

```
# fluvax_map$YEAR


#now take >6 months, which is vaccination in all ages. Do 5-12 separately
flu_vax_filt_allages <- flu_vax_filt_2 |> filter(Dimension == "≥6 Months")

#get the cohort membership by state from the original dataframe
coh_fluvax_allages <- aggregate(NMEMB ~ STATE + YEAR, data = coh, sum)
flu_vax_rescale_allages <- left_join(flu_vax_filt_allages, coh_fluvax_allages) |>
  mutate(sample_vaxxed = as.numeric(`Estimate (%)`)*1/100*NMEMB)
#aggregate the members and sample vaxxed
flu_vax_rescale_allages_memb <- aggregate(NMEMB ~ part + YEAR + Month, data = flu_vax_rescale_allages, su
m)
flu_vax_rescale_allages_vax <- aggregate(sample_vaxxed ~ part + YEAR + Month, data = flu_vax_rescale_alla
ges, sum)
flu_vax_rescale_allages_comb <- left_join(flu_vax_rescale_allages_vax, flu_vax_rescale_allages_memb) |> m
utate(flu_coverage = sample_vaxxed / NMEMB* 100)


#aggregate(NMEMB ~ STATE + YEAR, data = by_month_regions, sum)
#aggregate(NMEMB ~ STATE + YEAR, data = flu_by_month_regions, sum)
#these are different by a factor of 12, which makes sense because in one case I am summing over all the m
onths.
#I think this is still what I want, but need to check with stephen. Shouldn't affect scaling though cuz i
ts just off by a factor of 12 which will come out regardless

#do the same thing for the 5-9 age group
flu_vax_filt_512 <- flu_vax_filt_2 |> filter(Dimension == "5-12 Years")
coh_fluvax_59 <- aggregate(NMEMB~ STATE + YEAR + AGEGRP, data = coh, sum) |> filter(AGEGRP == "05_09")
flu_vax_rescale_5 <- left_join(flu_vax_filt_512, coh_fluvax_59) |> mutate(sample_vaxxed = as.numeric(`Est
imate (%)`)*1/100*NMEMB)
flu_vax_rescale_5_memb <- aggregate(NMEMB ~ part + YEAR + Month, data = flu_vax_rescale_5, sum)
flu_vax_rescale_5_vax <- aggregate(sample_vaxxed ~ part + YEAR + Month, data = flu_vax_rescale_5, sum)
flu_vax_rescale_5_comb <- left_join(flu_vax_rescale_5_memb, flu_vax_rescale_5_vax) |>
  mutate(flu_coverage = sample_vaxxed / NMEMB * 100)

#plot the strep and flu vaccine data ontop of one another for all ages, and for 5-9 age group specificall
y
flu_vax_rescale_allages_comb <- flu_vax_rescale_allages_comb |> mutate(MONTH = Month) |> select(-Month)
joint2 <- by_month_regions_only |> select(-NVISITS, -NMEMB) |> left_join(flu_vax_rescale_allages_comb) |>
  select(-sample_vaxxed, -NMEMB) |> mutate(Strep_CI_per_ten_thousand = CI_per_thousand*10)
joint2_pivot <- joint2|> pivot_longer(cols = c("Strep_CI_per_ten_thousand", "flu_coverage"))
fluvax_strep_allages_plot <- joint2_pivot |> ggplot(aes(x= MONTH, y = value, color = part)) +
  geom_line(aes(linetype = name)) +
  facet_wrap(~YEAR) + scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG",
"SEPT",

"OCT", "NOV", "DEC")) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + ggtitle("St
rep Pharyngitis Rates and Flu Vaccination Over Time  \n All Ages, scaled to MarketScan maybe") + labs(lin
etype = "Value Plotted") + labs(color = "Region") + scale_linetype_discrete(labels = c("Flu Vaccination
%", "Strep cases per 10,000"))


#for 5-9 age group
by_month_regions_5 <- by_month_regions |> filter(AGEGRP %in% c("05_09"))
by_month_regions_5_vis <- aggregate(NVISITS ~ part  + YEAR + MONTH,
                                       dat = by_month_regions_5,
                                       sum)
by_month_regions_5_memb <- aggregate(NMEMB ~ part  + YEAR + MONTH,
                                       dat = by_month_regions_5,
                                       sum)
by_month_regions_5 <- left_join(by_month_regions_5_vis, by_month_regions_5_memb)
by_month_regions_5 <- by_month_regions_5 |> mutate("CI_per_thousand" = NVISITS/NMEMB *1000)
#fluvax data
flu_vax_rescale_5_comb <- flu_vax_rescale_5_comb |> mutate(MONTH = Month) |> select(-Month)
joint3 <- by_month_regions_5 |> select(-NVISITS, -NMEMB) |> left_join(flu_vax_rescale_5_comb) |>
  select(-sample_vaxxed, -NMEMB) |> mutate(Strep_CI_per_ten_thousand = CI_per_thousand*10, Strep_CI_per_f
ive_thousand = CI_per_thousand*5, Strep_CI_per_two_thousand = CI_per_thousand*2)
joint3_pivot <- joint3 |> pivot_longer(cols = c("Strep_CI_per_two_thousand", "flu_coverage"))
fluvax_strep_59_plot <- joint3_pivot |> ggplot(aes(x= MONTH, y = value, color = part)) +
  geom_line(aes(linetype = name)) +
  facet_wrap(~YEAR) + scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG",
```

```
"SEPT",

"OCT", "NOV", "DEC")) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + ggtitle("St
rep Pharyngitis Rates and Flu Vaccination Over Time  \n 5-9 Age Group") + labs(linetype = "Value Plotte
d") + labs(color = "Region") + scale_linetype_discrete(labels = c("Flu Vaccination %", "Strep cases per
2,000"))

# fluvax_strep_allages_plot
# fluvax_strep_59_plot
childvax_data <- read_csv("/Users/madeleinekline/Dropbox (Harvard University)/G1/GradLab/vax_data/Vaccina
tion_Coverage_among_Young_Children__0-35_Months_.csv")
pcv_vax <- childvax_data |> filter(Vaccine == "PCV")
pcv_4d_35m <- pcv_vax |> filter(Dose == "≥4 Doses") |> filter(Dimension == "35 Months")
pcv_4d_35m <- pcv_4d_35m |> filter(`Birth Year/Birth Cohort` %in% c('2011', '2012', '2013',
                                                    '2014', '2015', '2016', '2017', '2018')) |>
  mutate(year = as.numeric(`Birth Year/Birth Cohort`), region = tolower(Geography))
pcv_map <- pcv_4d_35m |> filter(year %in% c(2011,2012,2013)) |> left_join(us_df)
# pcv_over_time_4d35m <- us_base + geom_polygon(data = pcv_map, aes(fill = `Estimate (%)`))+
#   geom_polygon(color = "gray", fill = NA) + theme(
#   axis.text = element_blank(),
#   axis.line = element_blank(),
#   axis.ticks = element_blank(),
#   panel.border = element_blank(),
#   panel.grid = element_blank(),
#   axis.title = element_blank()
#   ) + facet_wrap(~year) + scale_fill_gradient(low = "#FFFFFF", high = "#800000" ) + ggtitle("PCV Vaccin
ation by 35 months by birth cohort")
# pcv_over_time_4d35m


pcv_grouping <- pcv_4d_35m |> filter(`Birth Year/Birth Cohort` %in% c("2011", "2012", "2013")) |>
  mutate(STATE = Geography, YEAR = as.numeric(`Birth Year/Birth Cohort`)) |>
  select(Vaccine, STATE, `Estimate (%)`, YEAR, `Sample Size`)
pcv_memb_ag <- coh |> filter(YEAR %in% c("2011", "2012", "2013"), AGEGRP == "00_04")
pcv_memb_ag <- aggregate(NMEMB ~ STATE + AGEGRP + YEAR, dat = pcv_memb_ag, sum)
pcv_memb_ag <- pcv_memb_ag |> mutate(YEAR = as.numeric(YEAR))
pcv_grouped <- left_join(pcv_grouping, pcv_memb_ag)
pcv_grouped <- pcv_grouped |> mutate(sample_vaxxed = `Estimate (%)`*0.01 *NMEMB)
pcv_grouped <- pcv_grouped |> mutate(state = tolower(STATE))
pcv_grouped <- left_join(pcv_grouped, state_parts_2)
pcv_grouped_2016 <- pcv_grouped |> filter(YEAR == 2011) |> na.omit() #NAs are for regions that aren't sta
tes; ok to omit
pcv_grouped_2017 <- pcv_grouped |> filter(YEAR %in% c(2011, 2012)) |> na.omit()
pcv_grouped_2018 <- pcv_grouped |> na.omit()

pcv_grouped_2016_vaxxed <- aggregate(sample_vaxxed ~ part, data = pcv_grouped_2016, sum)
pcv_grouped_2016_memb <- aggregate(NMEMB ~ part, data = pcv_grouped_2016, sum)
pcv_2016_ag_region <- left_join(pcv_grouped_2016_memb, pcv_grouped_2016_vaxxed) |> mutate(pcv_coverage =
sample_vaxxed / NMEMB * 100, YEAR = 2016)

pcv_grouped_2017_vaxxed <- aggregate(sample_vaxxed ~ part, data = pcv_grouped_2017, sum)
pcv_grouped_2017_memb <- aggregate(NMEMB ~part, data = pcv_grouped_2017, sum)
pcv_2017_ag_region <- left_join(pcv_grouped_2017_memb, pcv_grouped_2017_vaxxed) |> mutate(pcv_coverage =
sample_vaxxed / NMEMB * 100, YEAR = 2017)

pcv_grouped_2018_vaxxed <- aggregate(sample_vaxxed ~ part, data = pcv_grouped_2018, sum)
pcv_grouped_2018_memb <- aggregate(NMEMB ~part, data = pcv_grouped_2018, sum)
pcv_2018_ag_region <- left_join(pcv_grouped_2018_memb, pcv_grouped_2018_vaxxed) |> mutate(pcv_coverage =
sample_vaxxed / NMEMB * 100, YEAR = 2018)

all_pcv <- rbind(pcv_2016_ag_region, pcv_2017_ag_region, pcv_2018_ag_region) |> select(part, pcv_coverag
e, YEAR) #coverage for 5-9 year olds

#plot it ontop of strep cases
#look by region over time by age group
by_year_age_region_visits <- aggregate(NVISITS ~ part + AGEGRP + PRIMARYCOND + YEAR, dat = by_year_age, s
um)
by_year_age_region_members <- aggregate(NMEMB ~ part + AGEGRP + PRIMARYCOND + YEAR, dat = by_year_age, su
m)
by_year_age_region <- left_join(by_year_age_region_visits,by_year_age_region_members)
by_year_age_region <- by_year_age_region |> mutate(CI_per_hundred = NVISITS/NMEMB*100)
strep_for_pcv <- by_year_age_region |> filter(AGEGRP == "05_09", YEAR %in% c(2016, 2017, 2018)) |>
```

```
  mutate(Strep_CI_per_hundred = CI_per_hundred) |> select(part, YEAR, Strep_CI_per_hundred)
strep_pcv <- left_join(strep_for_pcv, all_pcv) |> mutate(Strep_CI_per_five_hundred = Strep_CI_per_hundred
*5) |>
  select(-Strep_CI_per_hundred)
strep_pcv_pivoted <- pivot_longer(strep_pcv, cols = c("Strep_CI_per_five_hundred", "pcv_coverage"))
# strep_pcv_pivoted|>
#   ggplot(aes(x= YEAR, y = value, color = part)) +
#   geom_line(aes(linetype = name)) + scale_x_discrete(limits = c(2016, 2017, 2018))

#maybe separate by region, do barplots
midwest <- strep_pcv_pivoted |> filter(part == "midwest") |> ggplot(aes(x = YEAR, y = value, fill = nam
e)) + geom_bar(stat = "identity", color = "black", position = position_dodge()) + labs(fill = "Value") +
scale_fill_manual(values = c("#800000", "#C21807" ), labels = c("PCV Vaccination %", "Strep Cases per 50
0")) + ylim(0,100) + ggtitle("PCV Coverage and Strep Pharyngitis Cases Midwest")

northeast <- strep_pcv_pivoted |> filter(part == "northeast") |> ggplot(aes(x = YEAR, y = value, fill = n
ame)) + geom_bar(stat = "identity", color = "black", position = position_dodge()) + labs(fill = "Value")
+ scale_fill_manual(values = c("#800000", "#658354" ), labels = c("PCV Vaccination %", "Strep Cases per 5
00")) + ylim(0,100) + ggtitle("PCV Coverage and Strep Pharyngitis Cases Northeast")

west <- strep_pcv_pivoted |> filter(part == "west") |> ggplot(aes(x = YEAR, y = value, fill = name)) + ge
om_bar(stat = "identity", color = "black", position = position_dodge()) + labs(fill = "Value") + scale_fi
ll_manual(values = c("#800000", "#7852A9" ), labels = c("PCV Vaccination %", "Strep Cases per 500")) + yl
im(0,100) + ggtitle("PCV Coverage and Strep Pharyngitis Cases West")

south <- strep_pcv_pivoted |> filter(part == "south") |> ggplot(aes(x = YEAR, y = value, fill = name)) +
geom_bar(stat = "identity", color = "black", position = position_dodge()) + labs(fill = "Value") + scale_
fill_manual(values = c("#800000", "#6699CC" ), labels = c("PCV Vaccination %", "Strep Cases per 500")) +
ylim(0,100) + ggtitle("PCV Coverage and Strep Pharyngitis Cases South")

# grid.arrange(midwest, northeast, west, south, ncol = 2, nrow =2)

joint4 <- flu_by_month_regions_only |> select(-NVISITS, -NMEMB, -CI_per_ten_thousand, -CI_per_hundred_tho
usand) |> mutate(Flu_CI_per_thousand = CI_per_thousand) |> select(-CI_per_thousand)
strep_flu_vax_allages <- left_join(joint2,joint4) |> mutate(Strep_CI_per_thousand = CI_per_thousand) |> s
elect(-CI_per_thousand, -Strep_CI_per_ten_thousand)

linmod1 <- lm(Strep_CI_per_thousand ~ part,strep_flu_vax_allages)
summary(linmod1)
strep_flu_vax_allages |> mutate(pred = predict(linmod1, strep_flu_vax_allages)) |> summarize(RMSE = sqrt
(mean(pred - Strep_CI_per_thousand)^2))
adjr2_1 <- summary(linmod1)$adj.r.squared
#RMSE 9.969183e-16, R^2: 0.3375, adjusted R^2 0.3328
#try adding in covariates
aic1 <- summary(glm(Strep_CI_per_thousand ~ part,strep_flu_vax_allages, family = gaussian))$aic
#AIC is 1032.722


#add year
linmod2 <- lm(Strep_CI_per_thousand ~ part + as.factor(YEAR), strep_flu_vax_allages)
summary(linmod2) #RMSE 1.040379e-15; R^2 0.3955, Adjusted R^2 0.3797
strep_flu_vax_allages |> mutate(pred = predict(linmod2, strep_flu_vax_allages)) |> summarize(RMSE = sqrt
(mean(pred - Strep_CI_per_thousand)^2))
aic2 <- summary(glm(Strep_CI_per_thousand ~ part + as.factor(YEAR),strep_flu_vax_allages, family = gaussi
an))$aic
#AIC 1009.125
adjr2_2 <- summary(linmod2)$adj.r.squared

#add month
linmod3 <- lm(Strep_CI_per_thousand ~ part + as.factor(YEAR) + as.factor(MONTH),strep_flu_vax_allages)
summary(linmod3) #RMSE 1.345368e-15  R^2 0.865, adjusted R^2 0.8578
strep_flu_vax_allages |> mutate(pred = predict(linmod3, strep_flu_vax_allages)) |> summarize(RMSE = sqrt
(mean(pred - Strep_CI_per_thousand)^2))
adjr2_3 <- summary(linmod3)$adj.r.squared
aic3 <- summary(glm(Strep_CI_per_thousand ~ part + as.factor(YEAR) + as.factor(MONTH),strep_flu_vax_allag
es, family = gaussian))$aic
#AIC 383.3943

#adding in flu vaccination coverage
linmod4 <- lm(Strep_CI_per_thousand ~ part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage, strep_flu
_vax_allages)
summary(linmod4) #RMSE 2.252061e-15   R^2 0.873, adjusted R^2 0.8654
```

```
strep_flu_vax_allages_nojune <- strep_flu_vax_allages |> filter(MONTH != 6)
adjr2_4 <- summary(linmod4)$adj.r.squared
strep_flu_vax_allages_nojune |> mutate(pred = predict(linmod4, strep_flu_vax_allages_nojune)) |> summariz
e(RMSE = sqrt(mean(pred - Strep_CI_per_thousand, na.rm = TRUE)^2))
aic4 <- summary(glm(Strep_CI_per_thousand ~ part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage,stre
p_flu_vax_allages, family = gaussian))$aic
#AIC 339.7997

#adding in flu CI as well
linmod5 <- lm(Strep_CI_per_thousand ~ part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage + Flu_CI_p
er_thousand,strep_flu_vax_allages)
summary(linmod5) #RMSE 1.86077e-15 R^2 0.8878, adjusted R^2 0.8807
adjr2_5 <- summary(linmod5)$adj.r.squared
strep_flu_vax_allages_nojune |> mutate(pred = predict(linmod5, strep_flu_vax_allages_nojune)) |> summariz
e(RMSE = sqrt(mean(pred - Strep_CI_per_thousand, na.rm = TRUE)^2))
aic5 <- summary(glm(Strep_CI_per_thousand ~ part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage + Fl
u_CI_per_thousand,strep_flu_vax_allages, family = gaussian))$aic
#AIC 293.772

#try a month and year interaction term
linmod6 <- lm(Strep_CI_per_thousand ~ part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage + Flu_CI_p
er_thousand + as.factor(YEAR)*as.factor(MONTH),strep_flu_vax_allages)
summary(linmod6) #R^2 0.9191, adjusted R-squared 0.8905
adjr2_6 <- summary(linmod6)$adj.r.squared
aic6 <- summary(glm(Strep_CI_per_thousand ~ part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage + Fl
u_CI_per_thousand + as.factor(YEAR)*as.factor(MONTH),strep_flu_vax_allages, family = gaussian))$aic
#AIC 323.0887

#trying without year
linmod7 <- lm(Strep_CI_per_thousand ~ part + as.factor(MONTH) + flu_coverage + Flu_CI_per_thousand,strep_
flu_vax_allages)
summary(linmod7) #R^2 0.8339, adjusted r^2 0.8272
adjr2_7 <- summary(linmod7)$adj.r.squared
aic7 <- summary(glm(Strep_CI_per_thousand ~ part + as.factor(MONTH) + flu_coverage + Flu_CI_per_thousand,
strep_flu_vax_allages, family = gaussian))$aic
#AIC 430.0374

#likelihood ratio test of full model with interactions vs without
lrtest(linmod5, linmod6) #p value of 0.0004; this thinks the interaction term is better

#make a plot of the aics
model_inclusion <- c("Region only", "Region and Year", "Region, Year, Month", "Region, Year, Month, Flu V
accination", "Region, Year, Month, Flu Vaccination, Flu Cases", "Region, Year, Month, Flu Vaccination, Fl
u Cases,Year and Month interaction", "Region, Month, Flu Coverage, Flu Cases")
AICs <- c(aic1, aic2, aic3, aic4, aic5, aic6, aic7)
R2s <- c(adjr2_1, adjr2_2, adjr2_3, adjr2_4, adjr2_5 ,adjr2_6, adjr2_7)


model_df <- data.frame(Models = model_inclusion, AIC = AICs, "R2" <- R2s)
R2_order <- model_df |> arrange(R2) |> pull(Models)
model_df |> ggplot(aes(x = factor(Models, level = R2_order ), y = R2)) +
  geom_point() +
  ylim(0,1) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1, size = 10), axis.title.x = element_b
lank()) +
  labs(title = "Adjusted R-squared by Model Parameter Inclusion")
level_order <- model_df |> arrange(desc(AIC)) |> pull(Models)
model_df |> ggplot(aes(x = factor(Models, level = level_order ), y = AIC)) +
  geom_point() +
  ylim(0,1100) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1, size = 10), axis.title.x = element_b
lank()) +
  labs(title = "AIC by Model Parameter Inclusion")

strep_flu_vax_allages_nojune |>
  mutate(pred = predict(linmod5, strep_flu_vax_allages_nojune)) |>
  pivot_longer(cols = c("Strep_CI_per_thousand", "pred")) |>
  ggplot(aes(MONTH, value, color = part)) +
  geom_line(aes(linetype = name)) +
  facet_wrap(~YEAR) +
```

```
  scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT", "OCT", "NO
V", "DEC")) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  scale_linetype_discrete(name= "Value Plotted", labels = c("Predicted Strep Cases per 1000 People", "Obs
erved Strep Cases per 1000 People")) + labs(color = "Region", title = "Predicted and Observed Streptococc
al Pharyngitis Cases, All Ages")
south <- strep_flu_vax_allages_nojune |>
  mutate(pred = predict(linmod5, strep_flu_vax_allages_nojune),
         lower = as.data.frame(predict(linmod5, strep_flu_vax_allages_nojune, interval = "confidence"))$l
wr,
         upper = as.data.frame(predict(linmod5, strep_flu_vax_allages_nojune, interval = "confidence"))$u
pr) |>
  filter(part == "south") |>
  ggplot(aes(MONTH, Strep_CI_per_thousand)) +
  geom_line(color = "cyan") +
  geom_ribbon(aes(ymin = lower, ymax = upper), fill = "cyan", alpha = 0.2)+
  facet_wrap(~YEAR) +
  scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT", "OCT", "NO
V", "DEC")) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + labs(y = "Predicted Strep Cases p
er 1000 People") + ggtitle("South")

midwest <- strep_flu_vax_allages_nojune |>
  mutate(pred = predict(linmod5, strep_flu_vax_allages_nojune),
         lower = as.data.frame(predict(linmod5, strep_flu_vax_allages_nojune, interval = "confidence"))$l
wr,
         upper = as.data.frame(predict(linmod5, strep_flu_vax_allages_nojune, interval = "confidence"))$u
pr) |>
  filter(part == "midwest") |>
  ggplot(aes(MONTH, Strep_CI_per_thousand)) +
  geom_line(color = "red") +
  geom_ribbon(aes(ymin = lower, ymax = upper), fill = "red", alpha = 0.2)+
  facet_wrap(~YEAR) +
  scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT", "OCT", "NO
V", "DEC")) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + labs(y = "Predicted Strep Cases p
er 1000 People") + ggtitle("Midwest")

northeast <-  strep_flu_vax_allages_nojune |>
  mutate(pred = predict(linmod5, strep_flu_vax_allages_nojune),
         lower = as.data.frame(predict(linmod5, strep_flu_vax_allages_nojune, interval = "confidence"))$l
wr,
         upper = as.data.frame(predict(linmod5, strep_flu_vax_allages_nojune, interval = "confidence"))$u
pr) |>
  filter(part == "northeast") |>
  ggplot(aes(MONTH, Strep_CI_per_thousand)) +
  geom_line(color = "green") +
  geom_ribbon(aes(ymin = lower, ymax = upper), fill = "green", alpha = 0.2)+
  facet_wrap(~YEAR) +
  scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT", "OCT", "NO
V", "DEC")) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + labs(y = "Predicted Strep Cases p
er 1000 People") + ggtitle("Northeast")

west <- strep_flu_vax_allages_nojune |>
  mutate(pred = predict(linmod5, strep_flu_vax_allages_nojune),
         lower = as.data.frame(predict(linmod5, strep_flu_vax_allages_nojune, interval = "confidence"))$l
wr,
         upper = as.data.frame(predict(linmod5, strep_flu_vax_allages_nojune, interval = "confidence"))$u
pr) |>
  filter(part == "west") |>
  ggplot(aes(MONTH, Strep_CI_per_thousand)) +
  geom_line(color = "purple") +
  geom_ribbon(aes(ymin = lower, ymax = upper), fill = "purple", alpha = 0.2)+
  facet_wrap(~YEAR) +
  scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT", "OCT", "NO
V", "DEC")) +
  theme_bw() +
```

```
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + labs(y = "Predicted Strep Cases p
er 1000 People") + ggtitle("West")


grid.arrange(south, west, northeast, midwest, ncol = 2)

#now try to fit the model just for those 3 years that has the PCV data as well, just for 5-9 years old
flu_by_month_regions_5 <- flu_by_month_regions |> filter(AGEGRP %in% c("05_09"))
flu_by_month_regions_5_vis <- aggregate(NVISITS ~ part  + YEAR + MONTH,
                                       dat = flu_by_month_regions_5,
                                       sum)
flu_by_month_regions_5_memb <- aggregate(NMEMB ~ part  + YEAR + MONTH,
                                       dat = flu_by_month_regions_5,
                                       sum)
flu_by_month_regions_5 <- left_join(flu_by_month_regions_5_vis, flu_by_month_regions_5_memb)
flu_by_month_regions_5 <- flu_by_month_regions_5 |>
  mutate("Flu_CI_per_thousand" = NVISITS/NMEMB *1000) |>
  select(-NVISITS, -NMEMB)
#need to get flu_vax data just for 5-9 age group
#joint3 has flu coverage and strep data
strep_flu_vax_pcv_59 <- left_join(joint3, flu_by_month_regions_5) |>
  left_join(strep_pcv) |>
  filter(YEAR %in% c(2016, 2017, 2018)) |>
  mutate(Strep_CI_per_thousand = CI_per_thousand) |>
  select(-CI_per_thousand, -Strep_CI_per_ten_thousand, -Strep_CI_per_five_thousand, -Strep_CI_per_two_tho
usand, -Strep_CI_per_five_hundred)
#do same model fit process
pcv_mod1 <- lm(Strep_CI_per_thousand ~part, strep_flu_vax_pcv_59)
summary(pcv_mod1) #R^2 0.3098, adjusted R^2 0.295
pr2_1 <- summary(pcv_mod1)$adj.r.squared
paic1 <- glm(Strep_CI_per_thousand ~part, strep_flu_vax_pcv_59, family = gaussian)$aic
#AIC 907.7759

#add year
pcv_mod2 <- lm(Strep_CI_per_thousand ~part + as.factor(YEAR), strep_flu_vax_pcv_59)
summary(pcv_mod2) #R^2 0.3169, adjusted R^20.2922
pr2_2 <- summary(pcv_mod2)$adj.r.squared
paic2 <- glm(Strep_CI_per_thousand ~part + as.factor(YEAR), strep_flu_vax_pcv_59, family = gaussian)$aic
#AIC 910.283

#add month
pcv_mod3 <- lm(Strep_CI_per_thousand ~part + as.factor(YEAR) + as.factor(MONTH), strep_flu_vax_pcv_59)
summary(pcv_mod3) #R^2 0.8539, adjusted R^2 0.8355
pr2_3 <- summary(pcv_mod3)$adj.r.squared
paic3 <- glm(Strep_CI_per_thousand ~part + as.factor(YEAR) + as.factor(MONTH), strep_flu_vax_pcv_59, fami
ly = gaussian)$aic
#AIC 710.1823


#add flu_coverage
pcv_mod4 <- lm(Strep_CI_per_thousand ~part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage, strep_flu
_vax_pcv_59)
summary(pcv_mod4) #R^2 is 0.8721, adjusted R^2 0.8544
pr2_4 <- summary(pcv_mod4)$adj.r.squared
paic4 <- glm(Strep_CI_per_thousand ~part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage, strep_flu_v
ax_pcv_59, family = gaussian)$aic
#AIC 641.2862

#add flu cases
pcv_mod5 <- lm(Strep_CI_per_thousand ~part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage + Flu_CI_p
er_thousand, strep_flu_vax_pcv_59)
summary(pcv_mod5) #R^2 0.8983, adjusted R^2 0.8831
pr2_5 <- summary(pcv_mod5)$adj.r.squared
paic5 <- glm(Strep_CI_per_thousand ~part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage + Flu_CI_per
_thousand, strep_flu_vax_pcv_59, family = gaussian)$aic
#AIC 613.0843


#add pcv coverage
pcv_mod6 <- lm(Strep_CI_per_thousand ~part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage + Flu_CI_p
er_thousand + pcv_coverage, strep_flu_vax_pcv_59)
summary(pcv_mod6) #R^2 0.8991, adjusted R^2 0.883; pretty much no change
```

```
pr2_6 <- summary(pcv_mod6)$adj.r.squared
paic6 <- glm(Strep_CI_per_thousand ~part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage + Flu_CI_per
_thousand + pcv_coverage, strep_flu_vax_pcv_59, family = gaussian)$aic
#AIC 614.0274

#likelihood ratio test of adding vs not adding PCV coverage
lrtest(pcv_mod5, pcv_mod6) #p-value is 0.3039; no benefit to adding PCV coverage

pcv_mod7 <- lm(Strep_CI_per_thousand ~part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage + Flu_CI_p
er_thousand + pcv_coverage + as.factor(YEAR)*as.factor(MONTH), strep_flu_vax_pcv_59)
summary(pcv_mod7) #R^2 0.9156, adjusted R^2 0.8811
pr2_7 <- summary(pcv_mod7)$adj.r.squared
paic7 <- glm(Strep_CI_per_thousand ~part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage + Flu_CI_per
_thousand + pcv_coverage + as.factor(YEAR)*as.factor(MONTH), strep_flu_vax_pcv_59, family = gaussian)$aic
#AIC 630.5204

pcv_mod8 <- lm(Strep_CI_per_thousand ~part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage + Flu_CI_p
er_thousand + as.factor(YEAR)*as.factor(MONTH), strep_flu_vax_pcv_59)
summary(pcv_mod8) #R^2 0.9143, adjusted R^2 0.8805
pr2_8 <- summary(pcv_mod8)$adj.r.squared
paic8 <- glm(Strep_CI_per_thousand ~part + as.factor(YEAR) + as.factor(MONTH) + flu_coverage + Flu_CI_per
_thousand + as.factor(YEAR)*as.factor(MONTH), strep_flu_vax_pcv_59, family = gaussian)$aic
#AIC 630.5139


lrtest(pcv_mod6, pcv_mod7) #p-value for interaction vs no interaction is 0.2646; not significant

pmodel_inclusion <- c("Region only", "Region and Year", "Region, Year, and Month","Region, Year, Month, a
nd Flu Vaccination", "Region, Year, Month, Flu Vaccination, and Flu Cases", "Region, Year, Month, Flu Vac
cination, Flu Cases, and PCV Vaccination", "Region, Year, Month, Flu Vaccination, Flu Cases, PCV Vaccinat
ion, Year*Month", "Region, Year, Month, Flu Vaccination, Flu Cases, Year*Month")
pAICs <- c(paic1, paic2, paic3, paic4, paic5, paic6, paic7, paic8)
pR2s <- c(pr2_1, pr2_2, pr2_3, pr2_4, pr2_5, pr2_6, pr2_7, pr2_8)

p_model_df <- data.frame(Models = pmodel_inclusion, AIC = pAICs, R2 = pR2s)
pR2_levels <- p_model_df |> arrange(R2) |> pull(Models)
p_level_order <- p_model_df |> arrange(desc(AIC)) |> pull(Models)

p_model_df |> ggplot(aes(x = factor(Models, level = pR2_levels ), y = R2)) +
  geom_point() +
  ylim(0,1) +
  theme_bw()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1, size = 10), axis.title.x = element_b
lank()) +
  labs(title = "Adjusted R-squared by Model Parameter Inclusion, 2016-2018, Ages 5-9")
p_model_df |> ggplot(aes(x = factor(Models, level = p_level_order ), y = AIC)) +
  geom_point() +
  ylim(0,950) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1, size = 10), axis.title.x = element_b
lank()) +
  labs(title = "AIC by Model Parameter Inclusion, 2016-2018, Ages 5-9")

strep_flu_vax_pcv_59_nojune <- strep_flu_vax_pcv_59 |>
  filter(MONTH != 6)
strep_flu_vax_pcv_59_nojune |>
  mutate(pred = predict(pcv_mod5, strep_flu_vax_pcv_59_nojune),
         lower = as.data.frame(predict(pcv_mod5, strep_flu_vax_pcv_59_nojune, interval = "confidence"))$l
wr,
         upper = as.data.frame(predict(pcv_mod5, strep_flu_vax_pcv_59_nojune, interval = "confidence"))$u
pr) |>
  pivot_longer(cols = c("Strep_CI_per_thousand", "pred")) |>
  ggplot(aes(MONTH, value, color = part)) +
  geom_line(aes(linetype = name)) +
  facet_wrap(~YEAR) +
  scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT", "OCT", "NO
V", "DEC")) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  scale_linetype_discrete(name= "Value Plotted", labels = c("Predicted Strep Cases per 1000 People", "Obs
erved Strep Cases per 1000 People")) + labs(color = "Region")
```

```
south <- strep_flu_vax_pcv_59_nojune |>
  mutate(pred = predict(pcv_mod5, strep_flu_vax_pcv_59_nojune),
         lower = as.data.frame(predict(pcv_mod5, strep_flu_vax_pcv_59_nojune, interval = "confidence"))$l
wr,
         upper = as.data.frame(predict(pcv_mod5, strep_flu_vax_pcv_59_nojune, interval = "confidence"))$u
pr) |>
  filter(part == "south") |>
  ggplot(aes(MONTH, Strep_CI_per_thousand)) +
  geom_line(color = "cyan") +
  geom_ribbon(aes(ymin = lower, ymax = upper), fill = "cyan", alpha = 0.2)+
  facet_wrap(~YEAR) +
  scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT", "OCT", "NO
V", "DEC")) +
  theme_bw() + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + labs(y = "Predicted
Strep Cases per 1000 People") + ggtitle("South")
midwest <- strep_flu_vax_pcv_59_nojune |>
  mutate(pred = predict(pcv_mod5, strep_flu_vax_pcv_59_nojune),
         lower = as.data.frame(predict(pcv_mod5, strep_flu_vax_pcv_59_nojune, interval = "confidence"))$l
wr,
         upper = as.data.frame(predict(pcv_mod5, strep_flu_vax_pcv_59_nojune, interval = "confidence"))$u
pr) |>
  filter(part == "midwest") |>
  ggplot(aes(MONTH, Strep_CI_per_thousand)) +
  geom_line(color = "red") +
  geom_ribbon(aes(ymin = lower, ymax = upper), fill = "red", alpha = 0.2)+
  facet_wrap(~YEAR) +
  scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT", "OCT", "NO
V", "DEC")) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + labs(y = "Predicted Strep Cases p
er 1000 People") + ggtitle("Midwest")
northeast <- strep_flu_vax_pcv_59_nojune |>
  mutate(pred = predict(pcv_mod5, strep_flu_vax_pcv_59_nojune),
         lower = as.data.frame(predict(pcv_mod5, strep_flu_vax_pcv_59_nojune, interval = "confidence"))$l
wr,
         upper = as.data.frame(predict(pcv_mod5, strep_flu_vax_pcv_59_nojune, interval = "confidence"))$u
pr) |>
  filter(part == "northeast") |>
  ggplot(aes(MONTH, Strep_CI_per_thousand)) +
  geom_line(color = "green") +
  geom_ribbon(aes(ymin = lower, ymax = upper), fill = "green", alpha = 0.2)+
  facet_wrap(~YEAR) +
  scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT", "OCT", "NO
V", "DEC")) +
  theme_bw() + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + labs(y = "Predicted
Strep Cases per 1000 People") + ggtitle("Northeast")
west <- strep_flu_vax_pcv_59_nojune |>
  mutate(pred = predict(pcv_mod5, strep_flu_vax_pcv_59_nojune),
         lower = as.data.frame(predict(pcv_mod5, strep_flu_vax_pcv_59_nojune, interval = "confidence"))$l
wr,
         upper = as.data.frame(predict(pcv_mod5, strep_flu_vax_pcv_59_nojune, interval = "confidence"))$u
pr) |>
  filter(part == "west") |>
  ggplot(aes(MONTH, Strep_CI_per_thousand)) +
  geom_line(color = "purple") +
  geom_ribbon(aes(ymin = lower, ymax = upper), fill = "purple", alpha = 0.2)+
  facet_wrap(~YEAR) +
  scale_x_discrete(limits = c("JAN", "FEB", "MAR", "APR", "MAY", "JUN", "JUL", "AUG", "SEPT", "OCT", "NO
V", "DEC")) +
  labs(y = "Predicted Strep Cases per 1000 People") + ggtitle("West") + theme_bw() + theme(axis.text.x =
element_text(angle = 90, vjust = 0.5, hjust=1))


grid.arrange(south, west, northeast, midwest, ncol = 2)
```

Conclusions

My analyses show regional trends in cases of streptococcal pharyngitis over the course of the year. In the West, cases of streptococcal pharyngitis are lower throughout the year than in the three other regions. In the South, cases are higher, and they are higher specifically starting at the end of the summer and into the early fall. Flu cases so similar but not identical trends; the South shows a similar pattern of increased divergence from the other regions starting at the end of the summer into the fall, but not to the same extent as the streptococcal pharyngitis cases. Additionally, the West does not show a similar lack of cases of flu; its pattern follows closely with the Northeast and Midwest. Flu cases and flu vaccination are mildly correlated with streptococcal pharyngitis cases, but PCV vaccination does not seem to trend with streptococcal pharyngitis cases. The linear model that I fit was able to predict trends reasonably well, although there is some room for improvement especially in the West.

Many open questions remain. Are the trends seen in this analysis unique to streptococcal pharyngitis? If so, what factors could explain it? Are there more specific states of geographic regions that are driving this trends? Given more time, it would be useful to account for other variables, including weather. It would also be useful to look at more diseases of the upper respiratory system, for example, RSV, parainfluenza, etc, as well as diseases that spread in a similar age group but are not respiratory, such as impetigo or gastroenteritis. Given that we've seen a massive change in the landscape of respiratory illnesses in the time since these data were collected with the emergence of COVID-19, and now a tri-demic of RSV, flu, and COVID-19, it would be interesting to see whether our understanding of the drivers of streptococcal pharyngitis continue to hold up in the present day.

references

1. Ellen R Wald. Group A streptococcal tonsillopharyngitis in children and adolescents: Clinical features and diagnosis. https://www.uptodate.com/contents/group-a-streptococcal-tonsillopharyngitis-in-children-and-adolescents-clinical-features-and-diagnosis/print?search=strep (https://www.uptodate.com/contents/group-a-streptococcal-tonsillopharyngitis-in-children-and-adolescents-clinical-features-and-diagnosis/print?search=strep).
2. Bisno, A. L. Acute Pharyngitis: Etiology and Diagnosis. Pediatrics 97, 949–954 (1996).
3. Luo, R. et al. Diagnosis and Management of Group a Streptococcal Pharyngitis in the United States, 2011–2015. BMC Infect Dis 19, 193 (2019).
4. Weekly U.S. Influenza Surveillance Report | CDC. https://www.cdc.gov/flu/weekly/index.htm (https://www.cdc.gov/flu/weekly/index.htm).
5. FluVaxView | FluVaxView | Seasonal Influenza (Flu) | CDC. https://www.cdc.gov/flu/fluvaxview/index.htm (https://www.cdc.gov/flu/fluvaxview/index.htm) (2022).
6. Kissler, S. M., Klevens, R. M., Barnett, M. L. & Grad, Y. H. Distinguishing the Roles of Antibiotic Stewardship and Reductions in Outpatient Visits in Generating a 5-Year Decline in Antibiotic Prescribing. Clinical Infectious Diseases 72, 1568–1576 (2021).
7. Ron Dagan, Stephen Pelton, Lauren Bakaletz, & Robert Cohen. Prevention of early episodes of otitis media by pneumococcal vaccines might reduce progression to complex disease. The Lancet Infectious Diseases 16,.
8. Ben-Shimol, S. et al. Impact of Widespread Introduction of Pneumococcal Conjugate Vaccines on Pneumococcal and Nonpneumococcal Otitis Media. Clinical Infectious Diseases 63, 611–618 (2016).
9. ChildVaxView | CDC. https://www.cdc.gov/vaccines/imz-managers/coverage/childvaxview/index.html (https://www.cdc.gov/vaccines/imz-managers/coverage/childvaxview/index.html) (2022).
10. Pneumococcal Vaccine Recommendations | CDC. https://www.cdc.gov/vaccines/vpd/pneumo/hcp/recommendations.html (https://www.cdc.gov/vaccines/vpd/pneumo/hcp/recommendations.html) (2022).

Thank you so much Stephen Kissler and Yonatan Grad for your help!