**Observational Studies: Influenza Analysis**

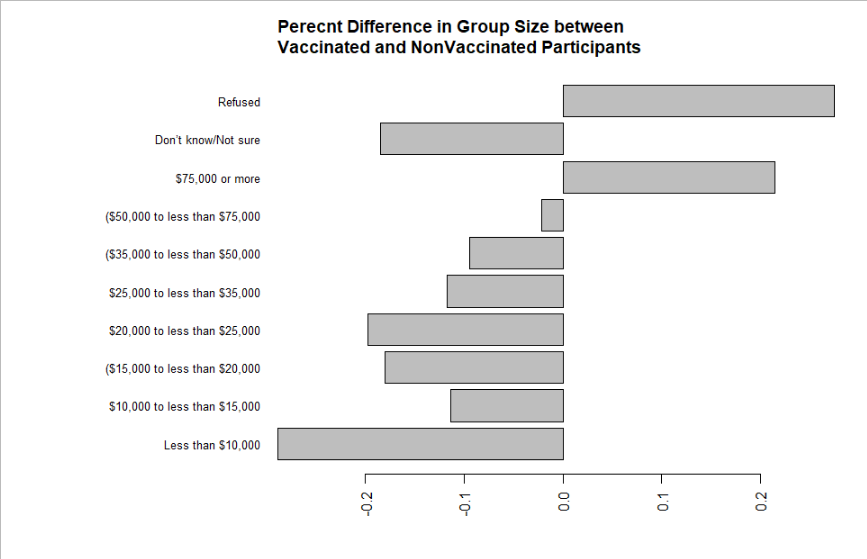
Christopher Barry

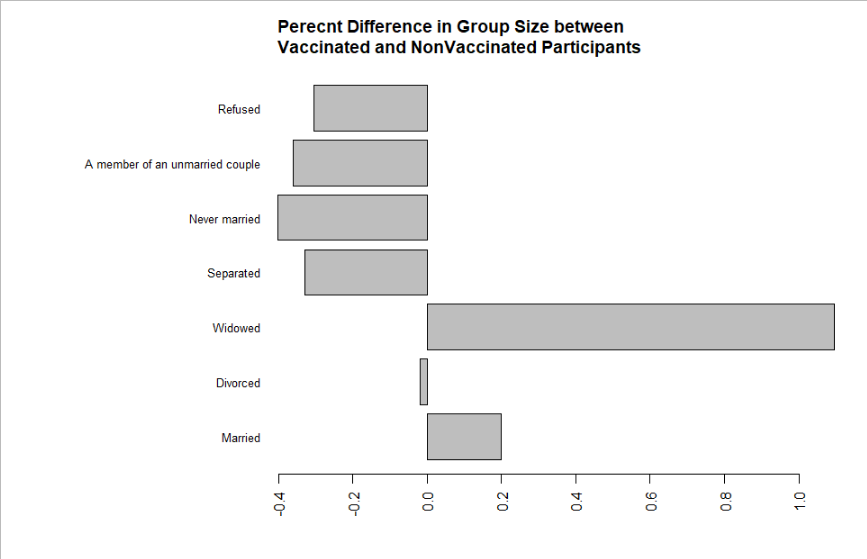
**Introduction:**

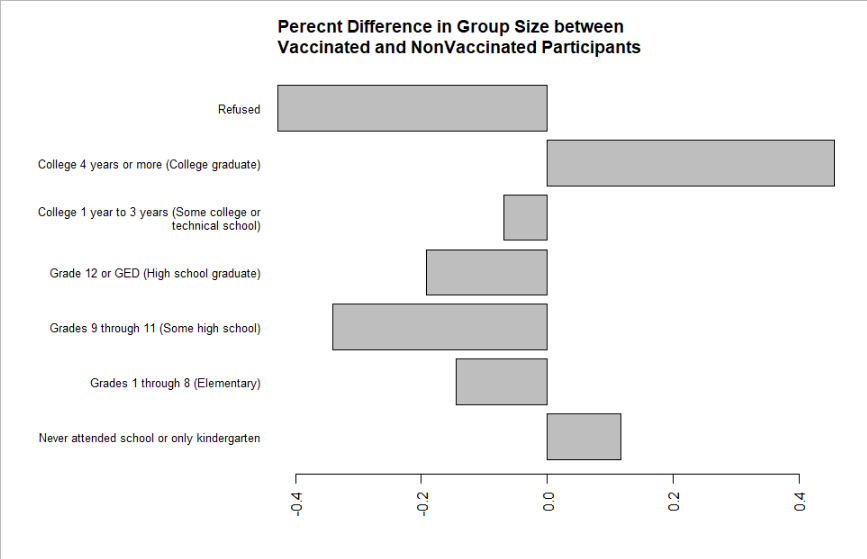
The influenza virus is a virus that most people on earth are familiar with, and is a virus that attacks the respiratory system of the host. On a national scale the flu affects many people: the Center for Disease Control (CDC) estimates that between 9.5 million and 45 million people have been infected with the influenza virus since 2010 *(CDC)*. When a person is infected this may result in inability to work and in some severe cases hospitalization and even death. During the months of December extending to March, there is a spike in United States influenza cases and is commonly referred to as flu season *(CDC)*. In order to prevent contracting the influenza virus health professionals at the CDC recommend that people practice proper hygiene, avoid close contact with others who are sick, and avoid touching your eyes, nose, and mouth. *(CDC)* Another prevention method for the influenza virus is vaccination. The influenza virus evolves every year and requires an annual vaccination. Although the vaccination is not guaranteed to work, it reduces the likelihood of contraction, and also can reduce the severity of a flu case.

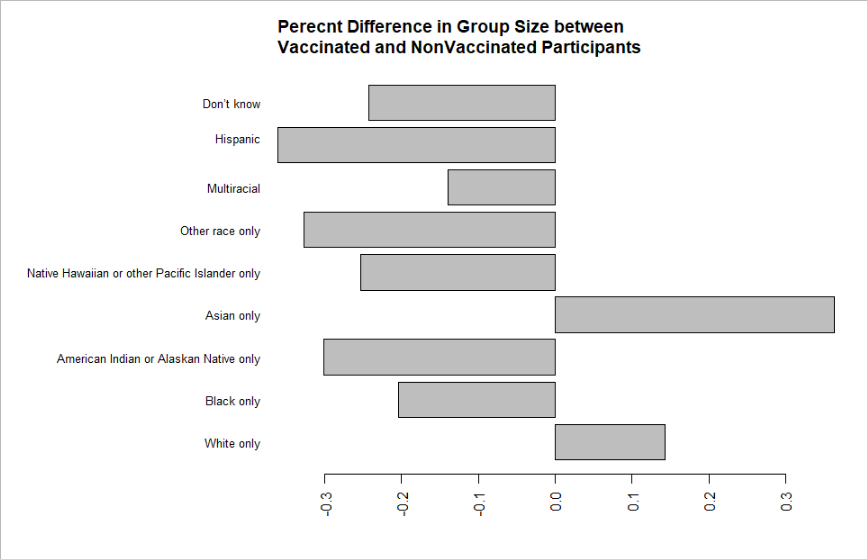
**Behavioral Risk Factor Surveillance System Data**

Every year the CDC conducts a survey called the Behavioral Risk Factor SUrveillance System (BRFSS). This survey is conducted over the phone with a random sample of the population, and is designed to give the CDC an idea of what health behaviors Americans are practicing. One of the questions is whether or not a person was vaccinated for the flu. The weighted percentage of survey participants responses for this each response for this question was 33.07% of respondents had received their influenza vaccination, and 66.42% had not received a vaccination. This number question is quite skewed in the direction of participants not receiving their flu shot. This in itself is quite alarming if we believe this sample of the population to be representative. It seems that many Americans do not receive a flu shot.

When observing flu vaccination through the lens of income, there seems to be an effect on flu vaccination. By using the svymean() and subset() functions, the percent proportion of income levels can be obtained for each income level for people who received a vaccination, and people who did not receive a vaccination. Next, we can take the percent difference between these two quantities and find out if there are any large differences between the makeup of the people who receive flu shots and people who do not receive flu shots by income level. The result from this is that all of the income classifications that are less than $50,000 are between 7.7% to 21.08% lower than the vaccinated group. For the group $50,000 to $75,000, the percent difference of participants belonging to this group when comparing vaccinations and non vaccinations is relatively flat in comparison with the lower income group (around 1%). For $75,000 and above, this group is 20% higher and the only meaningful group that is higher when comparing vaccinations and non vaccinations. This data tells a story that people who have a lower income up to a certain threshold (around $50,000), could be much less likely to get flu shots, where higher income levels are much more likely to obtain a flu shot vaccination. It may be worth investigating why people at lower income levels are not receiving as many flu shots, if it is lack of education of the impact of a flu shot, funds available, access to care, and availability of care. 

When observing vaccination through the lens of marital status, there again seems to be an effect that occurs. The percent of people who are married is 20% higher in the vaccinated group versus the non vaccinated group. Widowed participants in the vaccinated group are 100% higher than in the unvaccinated group. People who are separated, unmarried, and a member of an unmarried couple are all significantly lower between 31-39 percent lower. There are no large differences for divorced participants. These differences point to a direction the people who have someone in their life as the married participants do, could explain the increase in size. Another reason is that married people are twice as likely to have insurance. Widows may have a much higher value on life and are significantly more precautionary when considering their health. 

The next cut of flu shot vaccinations versus non flu shot vaccinations that was observed was education level. In this survey there were 6 possible answers to education level. The first percent difference comparison between group size for flu shot vaccinations versus non vaccinations is kindergartners and no school. This percent difference is slightly positive, however there are only 600 participants here. Choices 2-5 are all negative which start at grade 1 all the way through to 3 years of college. Once we reach choice 6 which are the college educated participants the vaccinated group is much larger than the unvaccinated group by 41%. One potential reason for this is because people who do not have a college degree are more likely to be working at a firm that does not offer health insurance. Also people with a four year degree may be working at a firm that pays for its employees to be vaccinated to prevent sickness in the workplace. This immediately removes barriers of cost and accessibility. There also may be some effect at play, where preventionary care is not deemed necessary to someone who is working in a lower paying job, because the opportunity cost is higher for obtaining a vaccination by spending little available time, energy, and money. 

The final demographic cut observed in this exploratory data analysis is race and ethinc group. This demographic question consisted of 8 different races. The only positive percent differences stemming from this demographic cut were white and asian people 14% and 25 percent respectively. These percent differences for every single race included in this survey were quite high for both negative and positive directions. The range of percent differences between vaccinated and non vaccinated group size was always greater than +/- 10%. Black, hispanic, Native American and Alaska Native, Native Hawaiian and Pacific Islander, mutli-racial, and other race were all negative, ranging from negative -32% to -17%. There may be some effects here due to cultural values of preventative care, between races. Differences in income and education levels may also be at play within races. This will need to be explored more before inclusion into the final model as variables. 

**The Outcome:**

The outcome variable of interest was pulled from the CDC BRFSS as well and will be the number of days that a survey participant indicated they were sick in the past 30 days. Through a few subsets of the entire data set this could be a good proxy to measuring whether a person had the flu or not. The first subset is whether or not this person believes their general health is acceptable. This will be helpful in filtering out any chronic diseases and physical injuries. If they believe their health is excellent, very good, good, or fair, these participants were included in the subset. The next criterion for honing in on the flu was subsetting the months when flu season spikes in the United States, which is between December to March. The question of physical health posed to participants is whether or not their physical health (illness or injury) was not good. Hopefully through these two subsets we will contain the best candidates that could have meant they contracted the flu when they responded a specific amount of days they were sick or injured. These subsets still leave us with a large amount of data at the individual level. By subsetting, approximately 62% of the data that was initially collected will be removed from this analysis.

**Methods:**

*Bayesian Additive Regression Trees (BART)*

To estimate the outcome, of whether or not a participant was sick or not sick the modeling method of Bayesian Additive Regression Trees was employed. This method uses a conditional distribution of the treatment either flu shot vaccination or no flu shot vaccination, and the covariates, with an additive error term attached to predict the Y response values. This conditional distribution is dependent on an ignorability assumption that assumes values of Y are orthogonal to values of treatments given individual values of covariates.

BART begins by separating treatment and nontreatment groups. Once this is done it uses the covariates to establish cut points which are represented by the different branches of the trees. Once the cut points are established a final stepwise function to fit the data for different values of covariates. Within BART there is a regularization parameter that helps control 1 variable from contributing too much to the fit when trees are constructed. This regularization parameter also ensures there are not too many nodes in the tree to fit the model. In order to find the model which has the best fit, BART uses a Monte Carlo Markov Chain to search through different tree parameters and variance to minimize the residual variance. Once a model is fit for an individual tree, this model explanatory power is subtracted from the initial y value. This model generates residuals and the next model is fit on the residual values only, treating these residual values as the y value. Once all of the sub trees have been fit,they are added up to give the yhat prediction value from BART. The idea is that weak learners do not perform on their own but adding them up together can provide accurate estimates. The individual tree contributions are not able to be realized by the model because the joint probability remains fixed at the initial tree calculations, however the tree structure is subject to change when each subsequent tree is calculated.

**Results**

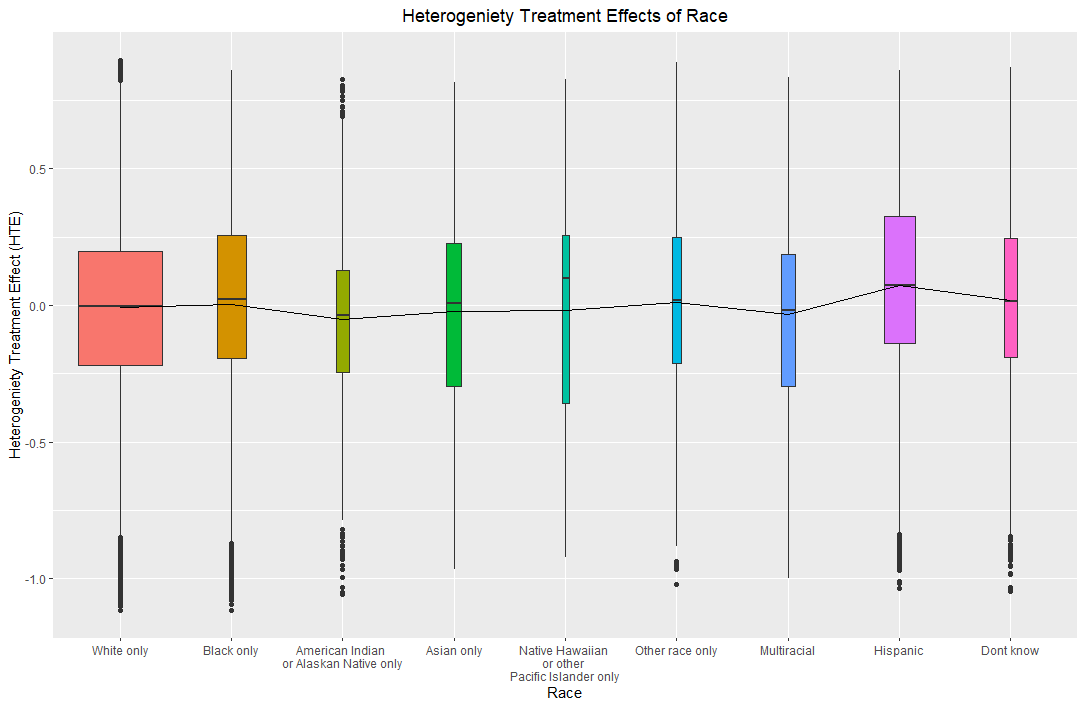
*BART*

The results that are output from BART include the train data predicted values and the test data predicted values for y. These values are calculated using a probit connection function because they are binary and this is the default function for BART when considering a binary outcome. Typically in medicine the group of interest is the experimental group or the treated group. Within this particular study, the treatment group is less interesting because there is fairly substantial belief in the medical community that flu shots are helpful preventative measures when attempting to lower a patient's chances of contracting the flu. Surprisingly the average treatment effect (ATE) is quite small, smaller than 0.01, between treated and non treated individuals in this study. This does not necessarily mean that flu shots do not work. With multiple covariates, they may have different treatment effects, exploring the average effect of treatment on the non-treated (ATNT), average treatment against the treatment (ATT) and the heterogeneity treatment effects (HTE) will give more insight into differences between covariates in this sample.

A potentially interesting method to view results is building the model, and constructing a holdout sample of the data, where the holdout dataset is two duplicate datasets of the observations that received either a treatment or no treatment stacked on top of each other. However the duplicate would have the opposite value assigned to the duplicate observations. This will allow the researcher to view the counterfactual prediction that the model would yield.

In this study, the ATT is small, because giving participants no flu shot who did receive a flu shot’s health outcome could easily be explained by other factors. People who receive flu shots could practice healthier lifestyles in general and live in communities of others who also receive flu shots. This would drastically reduce their risk of contracting the flu. The ATNT is significantly more interesting, because this taxonomy will vaccinate participants that did not receive a vaccination. However this treatment effect was found to be small, less than 0.01.

The next window to view the ATNT effects through is the heterogeneity treatment effect. This will take the ATNT from the non vaccinated observations that were duplicated and assigned a treatment, and subtract them from a non treatment. Then the individual differences are a participant level can be observed if they had or had not received a vaccination. All of the predicted values in this model were negative and in this probit model framework, a large absolute negative value is associated with a lower probability of sickness. For example if the no vaccination group is larger than the counterfactual, when the two values are subtracted from each other respectively a negative value will be obtained. A boxplot is a good tool for noting the differences between these differential values, and was used to identify any differences in distributions for covariates. For many covariates there is no discernible difference. The only difference that can be seen somewhat clearly is the race covariate. Below it can be seen that the boxplot for hispanic participants is higher than the others. Positive differences mean that counterfactual flu shot that was administered had a positive effect on health outcome. Another way to say this is that if a person who did not receive a flu shot, had actually received that flu shot they would have decreased their chances of contracting an illness. A potentially interesting suggestive finding of this study is that hispanic and black populations have a slightly positive skewed distribution. This means that these races may have benefitted from receiving a flu shot. The tree cut points help account for differences among the covariates, so these racial differences are independent of marital status, education level, and income. Potential sources of these differences could be racial biases that exist in the medical system or cultural characteristics that may influence less healthy lifestyles, or less vaccination. These are all difficult population characteristics to collect data on, and these findings remain solely speculative.



**Discussion**

The findings of this study do not show strong treatment effects in general, and for the covariates. This is contradictory to what medical research suggests and there are limitations that could explain why there are differences between this study and other work done in this field. The first assumption comes with the outcome of interest, this outcome is not a binary version of whether a participant contracted the flu, it is only have you been sick 5 days or more in the past 30 days, during the time period that flu season is considered active in the medical community. The possibility exists that many participants did not have the flu when they were sick for more than 5 days in the past month. It is probable that at least some of these participants who did get sick had the flu due to the widespread amount of cases in the United States population, however it is not possible to tell to what degree. These were big assumptions taken to construct the outcome.

The outcome and covariates are not independent of outside noise. If a flu shot helps reduce the possibility of contracting the virus for a subject, then it will reduce the likelihood of overall contractions. This is regardless of whether or not a participant has actually received a flu shot. Flu shots have become significantly more available to the public, where they can be received at most commercial drug stores for a low payment, and no insurance necessary. There is a possibility that enough people in communities are able to receive vaccinations and keep flu contraction fairly constant among different populations.

The slight uptick for black/hispanics that shows a benefit from giving unvaccinated black/hispanic populations the flu shot could point to a cultural difference. People who belong to these cultures could be less convinced of the science behind vaccinations and less likely to vaccinate. Additionally, if this effect does exist flu shot providers looking to make a profit are less likely to offer these vaccinations in areas where demand is low. There also could be discrimination against specific races within the healthcare system, and it could be harder for these minorty groups to obtain flu shots. There is evidence that discrimination exists but this data is difficult to collect measure and draw reasonable conclusions from.