BELLEVUE UNIVERSITY

What is the Efficacy of Flu Vaccines?

Business Problem

Flu season takes its toll on the world every year. With so many people getting ill or even dying, effective vaccination is critical in the fight against it. This project will examine a database containing results from several flu studies. The first step is to analyze the data and understand the different variables and what they each represent. Examination of the data leads to several business questions, with models constructed to look for significant correlations in the data. The project seeks to see if there are predictive measures in which donors will have a high or low response to specific influenza vaccinations. It further searches for indicators of who will respond favorably to a vaccination.

Background/History

The Flu, or Influenza, is a contagious respiratory illness caused by viruses that infect the nose, throat, and sometimes the lungs (Thomas, 2018). Every year, 5% to 20% of the U.S. population will get the Flu, with an average of 200,000 hospitalized with illness-related problems (DerSarkissian, 2017). 8,200 to 20,000 Americans die each year from flu-related causes. The average cost of hospitalizations and outpatient doctor visits related to the Flu is over $10 billion (DerSarkissian, 2017). Throughout history, there have been several flu pandemics that have made an impact worldwide. The 1889 Russian flu resulted in about 1 million deaths. The 1918 Spanish flu infected over half of the world's population, ending in over 40 to 50 million deaths. The 1957 Asian flu caused 1 million deaths, and the 1968 Hong Kong flu resulted in 1 to 3 million deaths. Most recently, the 2009 H1N1 flu led to 203,000 worldwide deaths (Vincent Iannelli, 2020). The Flu can be severe, but vaccines are available to protect against some strains. For example, in 2019, the "trivalent" flu vaccine protected against two Influenza A viruses and one influenza B virus (DerSarkissian, 2017).

Data Explanation

With the flu wreaking havoc every year, the importance of finding a more effective vaccine grows stronger. A research project investigating influenza vaccine imprints on the immune system, funded by the European Commission, led to the creation of the FluPRINT database.

This unified database results from a large-scale study exploring novel cellular and molecular underpinnings of successful immunity to influenza vaccines. This Database contains more than 3,000 parameters measured using various methods, including but not limited to mass cytometry, flow cytometry, phosphorylation-specific cytometry (phospho-flow), multiplex ELISA, clinical lab tests (hormones and complete blood. The dataset represents fully integrated and normalized immunology measurements from 747 individuals from eight clinical studies conducted between 2007 to 2015 at the Human Immune Monitoring Center of Stanford University. The dataset represents a unique source in terms of value and scale, which will broaden the understanding of influenza immunity (Andriana, Tomic, Dekker, Maecker, & Davis, 2019).

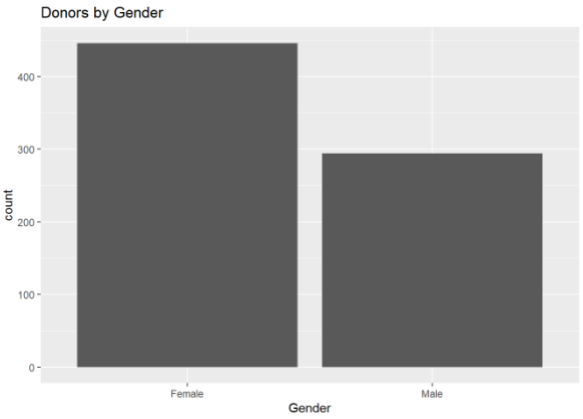
The data contains information regarding different vaccine types and the responses evoked in the donors. In addition, various tests were performed, indicating whether the vaccine's answer was high or low. Based on these tests, are there tests that are better indicators as to whether or not a donor will respond highly? In addition, are there any indicators that would predict whether a person will still get the Flu, even after vaccination?

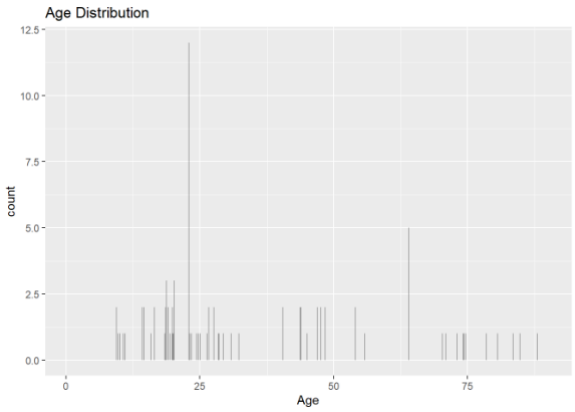
Methods

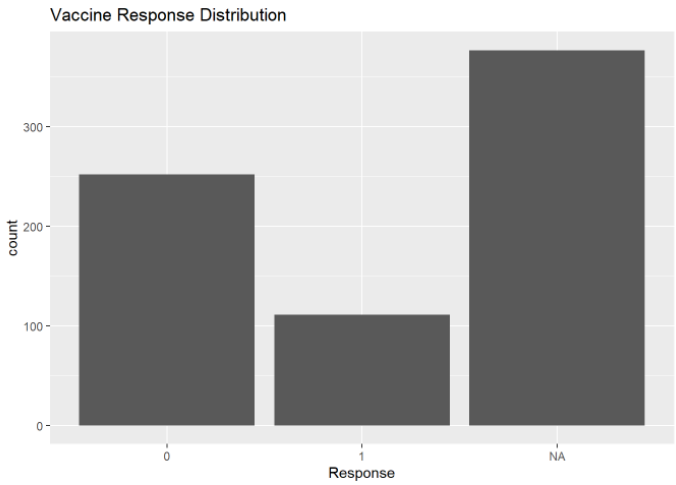
The data contains information on each donor, such as gender, race, and age. Additional information includes Body Mass Index (BMI), cytomegalovirus (CMV), Epstein-Barr virus (EBV), and the vaccine received at the visit. The vaccine outcome in the Database is recorded as a geometric mean titer (geomean). The difference in the geometric mean titers before and after vaccination (delta_geo_mean) and the difference for each vaccine strain (delta_single). The vaccine_resp field is recorded in the Database as a high or low responder. With each visit, samples were analyzed, and information about which assays were performed (assay field) and the value of the measured analytes (units and data) were recorded. The FluPRINT database recorded the donor's medical history, including past influenza vaccines over the last 5 years, influenza infection history, and influenza-related hospitalization (Andriana, Tomic, Dekker, Maecker, & Davis, 2019).

Exploratory data analysis was performed using Python and R software. The data was aggregated into another file that contained one record per donor to capture gender, race, age, cmv status, ebv status, bmi, statin use, vaccine type, vaccine response, and influenza history and hospitalization. Both of the files will be used for further regression testing. There are 740 donors, and the average age was 38. The youngest was 6 months old, while the oldest was 90. Donors by gender and age distributions are charted below:

The number of donors per response type is graphed here, with a 0 indicating a low response and 1 indicating a high response.







Analysis
Regression models were created using the aggregated donor file to test the correlations between several variables and outcomes. Using the aggregated dataset, I examined the influenza infection variable as a response to the type of vaccine, controlling for gender, age, and race of the donor, with no significant results. The following regression model tested the vaccine response as a response to the different vaccines administered, controlling for donor gender, age, and race. This showed a small significant correlation between vaccines 2 and 4 and a high response to the vaccine.

The following regression models were built using the entire dataset. First, a regression model was used to test the vaccine response variable as a response to the data variable, controlling for gender, age, and race of the donor. The data variable measured the values of the different assays tested on the donors. The model indicated a significant negative correlation between the data and the response to the vaccine. It also showed a positive correlation between the males and the age of the donors.

Because of this correlation, regression models were performed on subsets of the data on five different measured analytes with the most positive vaccine responses among donors (see Appendix A). These models indicated that the analytes measured for CD8+ T cells in the assay test CyTOF phenotyping had the most significant negative correlation. In addition, CD4+ T cells and T cells analytes had a minor significant positive correlation, and N.K. cells analytes had little negative considerable correlation. In contrast, B cells analytes had no significant correlation at all.

Conclusion
Based on the regression model performance, there is some significant correlation between vaccines 2 and 4 and having a high vaccine response. Vaccine 2 was the Fluzone Intradermal, and vaccine 4 was Fluzone. There seems to be some indication that the administration of these vaccines has a more substantial impact on a recipient having a better chance for increased response to influenza vaccination than the other types of vaccines.

The other result that seems to be of significance is the analyte used to measure CD8+ T cells. A low data score correlates significantly to high response to a vaccine. Therefore, this would seem to be a good indicator that a donor receiving a vaccine would have an improved chance of having a higher response to a vaccine if this measure has a low data value. This test could predict who will respond better to flu vaccinations. Determining who will respond favorably in the fight against the Flu could be an advantage.

Assumptions
As part of the vector selection and modeling, I have assumed Age, Gender, and Race as primary contributing factors to establishing the vaccine response based on the correlation analysis part of EDA. Based on this, regression models were built as well using a subset of data.

Limitations
The data set contained many inputs with missing variables, or input vectors deemed not completely relevant to the vaccine impact. Therefore, I had to clean the data to filter a few observations from the available data in the Database. Unfortunately, this leads to a reduction in the valuable data points, which might have reduced the regression model coverage.

Challenges
The data available in the dataset was collected using various tests, which resulted in attributes very specific to either one or another type of vaccine. This led to an explosion in input parameters with missing values across other types of vaccines. As a result, it needed proper cleaning to filter out unneeded observations.

Future Uses/Additional Applications
With more data, these models can be extended to validate the future efficacy of other types of vaccines.

Recommendations
Based on the data available, this model predicts that the efficacy of FlueZone vaccines is better with reasonable accuracy. But this model should be regressed again when more data is available.

Implementation Plan
With the Currently given parameters, this model can be launched to track the efficacy of all the different vaccine manufacturers. However, as mode data becomes available, the model must be redone to ensure there is no slippage due to data.

Ethical Assessment
Largely this model has been built based on the stand ord data set, with a limited data set. Therefore, there could be missing analysis or incorrect interpretation due to incomplete data collection for other vaccine manufacturers.
So, users of the model need to be careful in inferring outcomes and applying the actions in real-world scenarios.

References
Andriana, T., Tomic, I., Dekker, C. L., Maecker, H. T., & Davis, M. M. (2019). The FluPRINT dataset is a multidimensional analysis of the influenza vaccine imprint on the immune system. *BioRxiv*, Preprint https://doi.org/10.1101/564062.

DerSarkissian, C. (2017, November 16). *What Are Your Odds of Getting the Flu?* Retrieved from WebMD: https://www.webmd.com/cold-and-flu/flu-statistics

Gillespie, C. (2020, February 11). *This Is How Many People Die From the Flu Each Year, According to the CDC*. Retrieved from Health: https://www.health.com/condition/cold-flu-sinus/how-many-people-die-of-the-flu-every-year

Thomas, J. (2018, November 19). *The Flu: Facts, Statistics, and You*. Retrieved from Healthline: https://www.healthline.com/health/influenza/facts-and-statistics#1

Vincent Iannelli, M. (2020, February 5). *Annual Flu Deaths Among Adults and Children*. Retrieved from Very Well Health: https://www.verywellhealth.com/deaths-from-flu- 2633829

Appendix A

The names of analytes were counted based on having a high vaccine response. The top 5 are listed in the table below, along with their frequency counts. This was the basis for determining which analytes to conduct regression models.

CD8+ T cells	CD4+ T cells	T cells	B cells	NK cells	L50_CD40L
96	94	94	90	90	88

Possible Questions from the Audience

- What is the effect of Flu? Why do this project?
  Flu infects 5%-10% of the U.S population, causing around 20000 deaths. Most of the impacted are immunocompromised young or older adults. This infection cost roughly $10 Billion to the U.S. economy. Therefore, understanding which vaccine is effective is beneficial for saving people and reducing the impact on the U.S. economy.
- Why use FluPrint Database?
  FluPRINT Database is the largest collection of data collected from 740 donors during eight clinical experiments, with over 3000 parameters collected during these trials. This Database is extensive and covers the influenza history of each donor and the response to the vaccine. This helps to deduce the impacting factors for vaccine efficacy.
- Is the vaccine tests better indicators of whether a donor will respond highly?
  Yes, the regression analysis of the data shows that CD8+ T Cells are a significant indicator.
- Are there any indicators that would predict whether a person will still get the Flu, even after vaccination?
  There are some insignificant and significant correlators identified in part of the analysis that predicts this,
- What parameters did you consider using in your modeling?
  I have considered the donor's race, ethnicity, age, BMI, CMV infection, Epstein-Barr infection, and influenza history.
- What is the reason for selecting these specific parameters for your modeling?
  Based on the EDA, I have first established correlation, then chose the impacting factors.
- What is your method for preparing the data?
  I have used Python and R to aggregate the data and create a secondary input XLS for modeling.

- Are there any significant observations that can be interpreted from the data?
  Out of 740 Donors, about 60% are females. The age of Donors is spread between 6mo to 80 years. The largest concentration of donors is about 23 yr. old.
- What are the results from the modeling?
  In General, looks like there is some significance with Fluzone vaccines being protective.
- Are there any significant conclusions from this analysis? What are they?
  Based on the data, CD8+ T cells seem to be a significant indicator that if a person is vaccinated with a lower score vaccine, there is a higher chance of protection from the Flu Vaccine.
- Are there any ethical considerations in deploying this model into usage?
  Largely this model has been built based on the stand ord data set, with a limited data set. Therefore, there could be missing analysis or incorrect interpretation due to incomplete data collection for other vaccine manufacturers.
  So, users of the model need to be careful in inferring outcomes and applying the actions in real-world scenarios.