Learning from the Past

Assessing and Modeling Viral Vaccination in the Face of a Pandemic

Nicholas Garside—Mechanical Engineering PhD, 2nd year
Arinze Okafor—Biomedical Engineering PhD, 2nd year
Lucy Chikwetu—Biomedical Engineering PhD, 3rd year

Background

Throughout history, the outbreak of viral infections poses one of the greatest threats to human health and civilization. The occurrence of viral pandemics has, each time, resulted in the loss of thousands to millions of lives, the crippling of global economies and supply chain, and the straining of healthcare systems [4]. Vaccination remains one of the most effective long-term instruments to combat these debilitating effects. Despite its extreme importance, societal response to vaccination has not always been optimal and is influenced by several socioeconomic factors [6].

In 2009, an outbreak of a novel influenza A (H1N1 class) virus led to the death of over half a million people. Concerted efforts to tackle this outbreak led to the development of a new H1N1 viral vaccine only seven months after the initial outbreak. Despite this success, societal acceptance and vaccination rates for both the H1N1 virus and seasonal flu remained low [5] [1].

Following the path of history, the recent outbreak of the Severe Acute Respiratory Syndrome-Corona Virus 2 (SARS-CoV2) has resulted in the infection of over 30 million people worldwide, while causing the death of over a million people [7]. Despite this unsavory development, societal attitude to the ongoing global vaccination efforts remains less than optimal, with close to half of the US population having negative or pessimistic views [3]. Given this undesired societal attitude, even in the face of the current pandemic, and the importance of vaccination in saving lives and the global economy, it becomes imperative to utilize analytical tools to better understand the factors affecting vaccine acceptance, predict compliance to vaccination advisories, and guide the channeling of resources to maximize rates of vaccination in a bid to save lives.

Aim of Study

We hypothesize that the 2009 H1N1 viral outbreak allows us to learn about COVID-19 in retrospect. This study aims to harness available datasets from the 2009 pandemic and machine learning approaches in assessing factors affecting an individual's likelihood to get a vaccine. We plan to extrapolate these results to propose recommendations for implementing a future voluntary COVID-19 vaccine.

Data Description

The dataset came from an in-person National 2009 H1N1 Flue Survey conducted by the National Center for Health Statistics (NCHS) and the National Center for Immunization and Respiratory Diseases (NCIRD) in a bid to study the socio-economic effects influencing vaccine compliance. Nearly 40,000 United States citizens participated in the study. In addition to capturing demographic information, the survey also gauged individuals' opinions on H1N1 or seasonal flu vaccine effectiveness, solicited behavioral tendencies, and asked participants whether or not their doctor recommended an H1N1 or flu vaccine. The dataset has 35 features, a respondent ID, and two target variables—***h1n1_vaccine*** abd ***seasonal_vaccine***. Both target variables are dichotomous variables indicating whether or not an individual received a vaccine. The data includes a healthy mix of nominal variables, ordinal variables, and dichotomous features such as gender and marital status. Table 1. shows the summary of statistics for available features and the output.

Missing Values

Within the training dataset, there were 13201 observations (49.4%) with missing values in a total of 23 of the 35 predictor variables. Fig 2. is the visual representation of our missing values. Without a doubt, health insurance had the most missing values. We also recognize some unique patterns that facilitate smooth imputation. Those who did not fill out their opinions on particular issues tended not to fill out most of the opinion section. Children under six months old also missed information about whether or not they were healthcare workers, had health insurance, or had chronic medical conditions, and data imputation for such cases was easy. Fig 3. and Fig. 4 support the validity of those special cases. We also made intelligent inferences to impute missing values, for example, if an individual's employment status was listed as *Not in Labor Force* or *Unemployed* and they had no occupation, we created a new occupation and employent indusry factor called *Unemployed* and we would list the individual in that category.

Summary statistics for ordinal variables

	min	mean	median	max
h1n1_concern	0	1.618	2	3
h1n1_knowledge	0	1.263	1	2
opinion_h1n1_vacc_effective	1	3.851	4	5
opinion_h1n1_risk	1	2.343	2	5
opinion_h1n1_sick_from_vacc	1	2.358	2	5
opinion_seas_vacc_effective	1	4.026	4	5
opinion_seas_seas_risk	1	2.719	2	5
opinion_seas_sick_from_vacc	1	2.118	2	5
household_adults	0	0.887	1	3
household_children	0	0.535	0	3
age_group				
education				
income_poverty				

List of nominal variables

race, employment_status
employment_occupation
hhs_geo_region, census_msa, employment_industry

Summary statistics for dichotomous variables

	min	mean	median	max
behavioral_antiviral_meds	0	0.049	0	1
behavioral_avoidance	0	0.726	1	1
behavioral_face_mask	0	0.069	0	1
behavioral_wash_hands	0	0.826	1	1
behavioral_large_gatherings	0	0.359	0	1
behavioral_outside_home	0	0.337	0	1
behavioral_touch_face	0	0.677	1	1
doctor_recc_h1n1	0	0.220	0	1
doctor_recc_seasonal	0	0.330	0	1
chronic_med_condition	0	0.283	0	1
child_under_6_months	0	0.083	0	1
health_worker	0	0.1122	0	1
health_insurance	0	0.880	1	1
seasonal_vaccine	0	0.466	0	1
h1n1_vaccine	0	0.213	0	1
sex				
rent_or_own				
marital_status				

Table 1: Summary of statistics of all variables (including the output)

Distributions

We plotted stacked bar graphs to show each feature's relationship with the target variable—no strong associations where identified for the H1N1 vaccine, but interesting associations were identified for seasonal flu vaccine. H1N1 vaccine had a mean of 0.2125 and seasonal flu vaccine had 0.4656, and this could have contributed to the disparity we observed. Fig. 1 shows the plots for each feature's relationship with the seasonal flu vaccine.

Preliminary Relationships Between Features

The team did a preliminary analysis of the ordinal and binary features' correlations using Spearman's ranking method. Results from this analysis are shown in Fig. 5, with only significant correlations being displayed. This analysis revealed a significant positive correlation between the features: concern about the H1N1 virus, positive opinions on the vaccine, and age; and vaccine compliance ($p<0.05$). Also, older people seemed to have higher vaccination compliance ($p<0.05$), and people tend to comply more when their doctors recommend the vaccine ($p<0.05$). The relationship between the nominal variables and the outcome variables was accessed using the Chi-square test for statistical independence. This revealed that, of all six nominal variables (employment occupation, employment industry, employment occupation, region of residence as defined by the U.S. Dept. of Health and Human Services, region of residence by the metropolitan statistical areas and race), compliance with H1N1 vaccination were not independent of the respondent?s region of residence defined by the metropolitan areas (X-squared = 0.11411 , P=0.95). This was not the case for compliance with seasonal vaccination. However, this interpretation is limited as the metropolitan areas have been deidentified, making it difficult to know which areas were compliant and which were not.

Possible Modeling Approaches

This dataset is part of an online competition hosted by drivendata.org [2]. The competition's organizers split the dataset into two halves–one for training and another for testing. Label values (h1n1_vaccine and seasonal_vaccine) were only provided for the training data with a training data sample size of 26,707. The test dataset labels were withheld to facilitate the unbiased evaluation of model performance for the competition. We plan to further split the training set provided into training, validation, and testing datasets in order to enable model training, prevent model overfitting and enable model evaluation. We will split the data set into three groups using a split ratio of 60:20:20 for training, validation, and testing, respectively. To gain some insight into the fundamental linear and non-linear relationships between predictors and our outcome variable, we will explore less flexible but interpretable modeling approaches. Per our study goals and for the competition, we aim to develop a predictive model focusing on maximizing prediction accuracy. For this purpose, we will explore highly flexible approaches such as some tree-based methods (e.g., random forest, XGBoost), neural network-based methods (e.g., Multi-Layered Perceptron). Spanning the degree of model flexibility is a major goal, as we strive to both intuitively understand influencing factors and accurately predict outcomes as we apply our predictions to the real-world problem.

Conclusion

This project aims to answer the currently relevant question: What are the leading predictors of vaccination, and to what extent can we predict novel disease vaccination rates from these features? This question is becoming increasingly relevant as COVID-19 is nearing vaccine completion, for the answer could help guide public vaccination efforts in the future.

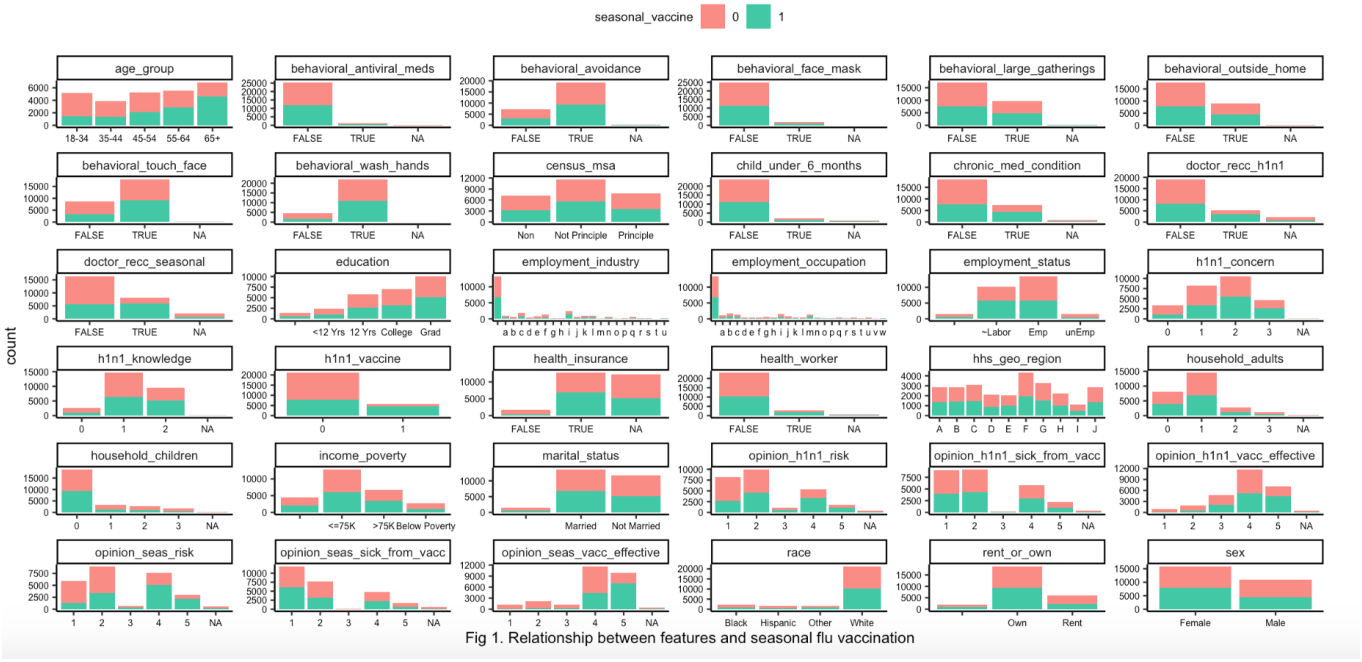

Fig 1. Relationship between features and seasonal flu vaccination

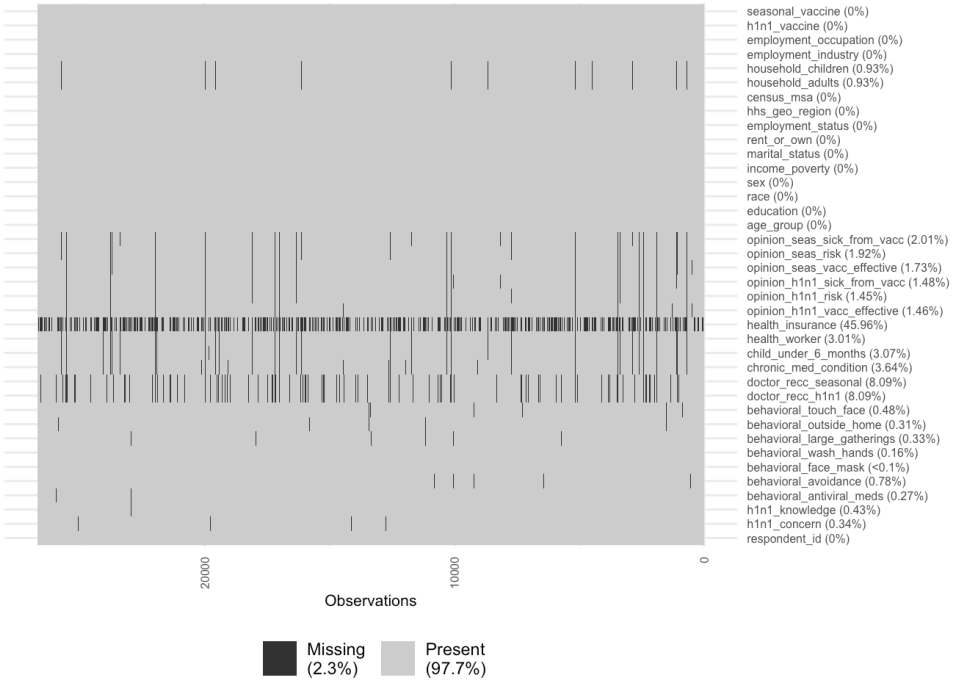

Fig 2. Visual representation of missing values

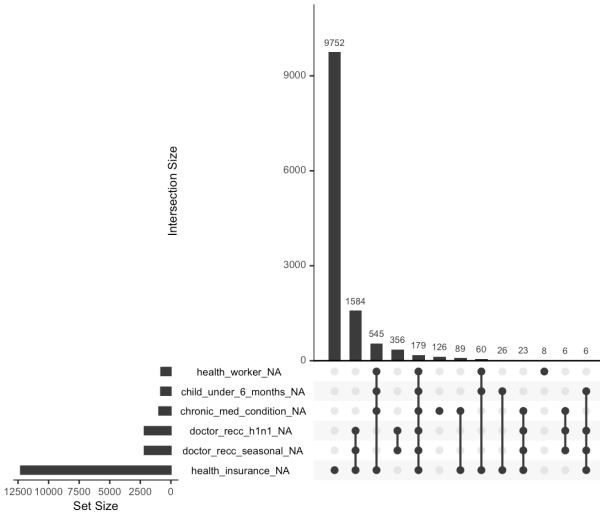

Fig 3. Most Missingness for Health Insurance is ungrouped

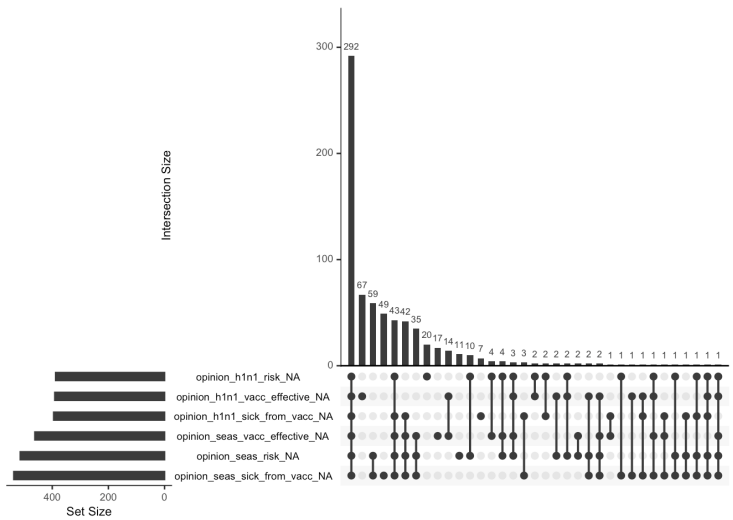

Fig 4. Missingness distributions for opinions

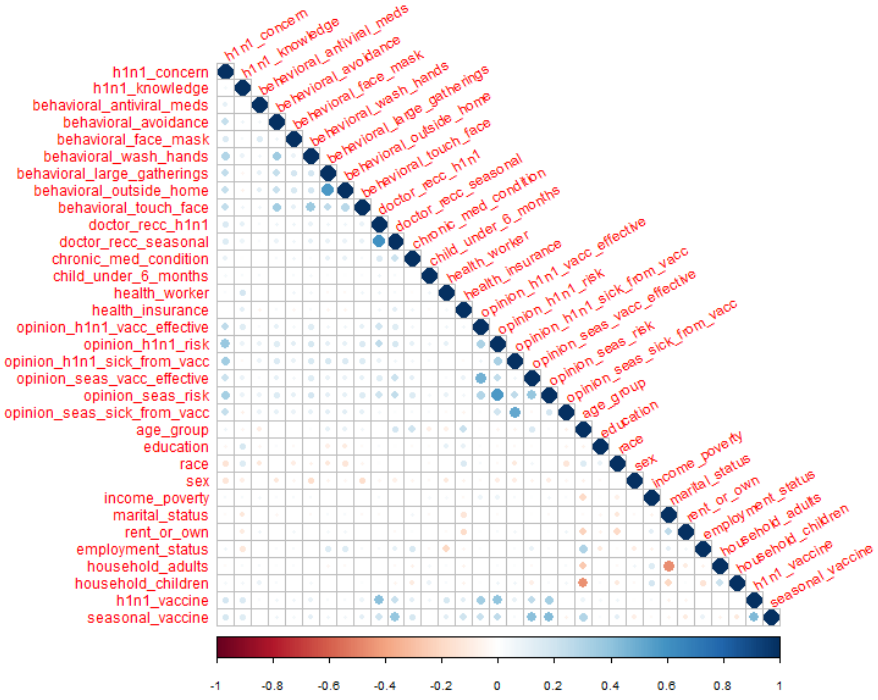

Fig. 5 Correlation plot for FluShot features

REFERENCES

[1] Final estimates for 2009–10 seasonal influenza and influenza a (H1N1) 2009 monovalent vaccination coverage ? united states, august 2009 through may, 2010., May 2011.

[2] DrivenData. Flu shot learning: Predict h1n1 and seasonal flu vaccines.

[3] Kimberly A. Fisher, Sarah J. Bloomstone, Jeremy Walder, Sybil Crawford, Hassan Fouayzi, and Kathleen M. Mazor. Attitudes toward a potential sars-cov-2 vaccine: A survey of u.s. adults. *Annals of Internal Medicine*, 2020.

[4] Damir Huremović. *Brief History of Pandemics (Pandemics Throughout History)*, pages 7–35. Springer International Publishing, 2019.

[5] Russell D. Ravert, Linda Y. Fu, and Gregory D. Zimet. Reasons for low pandemic h1n1 2009 vaccine acceptance within a college sample. *Advances in Preventive Medicine*, 2012:1?7, 2012.

[6] Louise E. Smith et al. A systematic review of factors affecting vaccine uptake in young children. *Vaccine*, 35(45):6059–6069, Oct 27 2017. Copyright - Copyright Elsevier Limited Oct 27, 2017; Last updated - 2019-07-22.

[7] Johns Hopkins University. Covid-19 dashboard by the center for systems science and engineering (CSSE) at johns hopkins university, 2020.