Socio-Demographic Charateristics Associated With H1N1 Vaccination Decision and Vaccine Predictions

ECO 475 Final Paper

Xinyi Qin, Ying Xiong, Yuxuan Wan

April 11, 2022

Keywords: Vaccination Prediction, H1N1, COVID-19, Logistic Regression, Random Forest

1. Introduction

1.1 Research Question

On 11th March 2020, the World Health Organization (WHO) officially declared the coronavirus outbreak a global pandemic. Statistics from the New York Times show that more than 80 million cases are tested positive in the United States, while nearly 1 million people have lost their lives by now. (New York Times, 2022) The coronavirus is a disease triggered by SARS-CoV-2, and it has evolved into multiple variants over time. Its prominent risks are developing respiratory illness, kidney failure, or, in the worst case, death. Depending on the cases, the symptoms normally show up within 2 to 14 days upon exposure to the virus, consisting of fever, shortness of breath, loss of taste and smell, body aches, etc. Nevertheless, those who have survived the fight against the virus still suffer from unfortunate outcomes that could potentially last a lifetime.

Many countries have implemented diverse guidance and rules, aiming to contain the increasing trend of positive cases and provide a robust and up-to-date response to the community. China, as an example, imposed restrictions on the mask mandates indoors and outdoors. Moreover, China has limited certain travelers' entry into the country, especially those with high body temperatures. People in some cities had to enter a multistage lockdown in which residents underwent a self-isolation period. The Wuhan lockdown is unprecedented in the public health history, but it is undeniable that the quarantine succeeded in reducing the spread of the virus and preventing a larger outbreak.

However, one primary consequence associated with the lockdown's restrictions is the inevitably sharp deceleration in economic growth. Due to the response to the coronavirus, the government implemented mandatory closure of nonessential businesses. For that reason, the U.S. economy shrunk immensely at an average annualized rate of 19.2%. Millions of people were forced to leave their occupations (Mutikani, 2021). The unemployment rate once soared

to almost 15% in May 2020 and has plummeted below 4% in 2022 (U.S. Department of Labo, 2022).

The top priority is to ease the economic contraction and control the outbreak through scientific methods. Thus, scientists and researchers devoted themselves to the vaccine development for the coronavirus at a rapid pace. The fastest vaccine that had previously been developed took four years, from viral sampling to approval (Philip, 2020). Since this pandemic is considered a severe global emergency, it took scientists roughly one year to accomplish it, with sufficient resources and funding. On 2nd December 2020, the United Kingdom approved the first vaccination for Covid-19. Soon, the vaccine rollout attracted enormous attention worldwide. After requiring individuals to take the second dose, two weeks following the first dose, the Pfizer vaccine is reported as 95% effective at averting the virus (Heidi et al., 2020). On 14th December, the appointments for the first dose of Covid-19 vaccines became available in 150 hospitals across the country for U.S. residents (BBC News, 2020). Nevertheless, less than 50% of the entire population was fully vaccinated by the end of June 2021.

The world has been staggered by the U.S.'s delayed response to the coronavirus since it has yielded the highest death counts. Particularly, we are interested in why people have procrastinated or refused to get their vaccination. If we can predict and identify the groups of people who are less likely to book their vaccine appointments, local health organizations can promote vaccines more effectively to the anticipated demographic. Since the up-to-date data related to the Covid-19 is not yet available, this paper will revisit the most recent respiratory disease pandemic, H1N1 influenza, in 2009, analyzing the survey data collected by the United States National Center for Health Statistics. Covid-19 and H1N1 share similarities as they are transmitted by physical contact with contaminated objects and can be spread asymptomatically. There is a significant overlap between H1N1 influenza and COVID-19 in terms of the

population at risk and resulting complications. Additionally, the effort on vaccine development was made quickly after the emergence of both diseases.

This paper will build logit regressions and machine learning models to examine the correlation between H1N1 vaccination decisions and social-demographic characteristics, including geography and economic background. By connecting the H1N1 vaccines and seasonal flu shots, we aim to further test both models' prediction power.

1.2 Literature Review

The emergence of H1N1 influenza has developed concerns from researchers, and, as a result, they have dug deeply into diverse aspects. Some relate to the origin of the virus and discuss the clinical pieces of evidence of transmission from the genomic point of view (Smith et al., 2009); some estimate the impact of social media and interpersonal interactions on the vaccination decisions (Carolyn, 2013); and others offer thoughtful insights on the current policy responses that best favor the politicians' and stakeholders' interest regarding the H1N1 vaccination (Baekkeskov, 2016).

In particular, numerous papers have indicated that the H1N1 virus, which occurred in March 2009, brought negative impacts on humans, yet it could be prevented and treated effectively through scientific methods, such as vaccine injections. As suggested by a clinical and epidemiological study, if the daily mean level of population immunity is improved by 10%, five cases of H1N1 can be prevented per day. Moreover, it is worth noting that without implementing mass vaccination programs against influenza, the number of cases that are tested positive under RIDT would be five-fold or more (Un-In et al., 2014). Moreover, a deficient level of actual vaccination was forecasted and observed in the absence of an increase in the risk perception of pandemic influenza H1N1 (Michel and Jocelyn, 2010). Therefore, specific policies and promotions have to be adequate to promote a wider vaccination uptake and further prevent a larger outbreak. In an article that discusses the perceptions and issues during the

H1N1 pandemic, the authors Henrich and Holmes (2011) examined the factors that influence the public's decisions on vaccine uptake by collecting comments on websites of primary news sources. They offered valuable insights by pointing out the key themes: the fear of H1N1 vaccines, doubtful attitudes towards government competency and trustworthiness, pharmaceutical companies, and personal protective measures. These factors have inspired us to question the fundamental reasons for not taking vaccines. Another article that examined the beliefs, attitudes, and practices associated with the intention to get vaccinated for the H1N1 virus within the general population in France offers a hint. The data acquired from the questionnaires included socio-demographic characteristics, risk and illness perceptions, and political attitudes. The result suggests that self-protective behavioral intentions tend to be relatively low in the general population, yet whether there is a clear relationship between the vaccination decision and the factors like geometric variable and risk perception remains unknown and requires further steps toward describing such relation.

To extend the literature, we use data from the National 2009 H1N1 Flu Survey (NHFS), which gathers information from households in all 52 states throughout the United States through computer-assisted random-digit-dialing telephone interviews. It is important to note that the group of interest is those who have their intentions and decisions to get vaccinated; hence, we excluded children under 18 years old. A variety of participant's characteristics like their risk perceptions towards the vaccine, state of residence, and age are adopted as the independent variables, and we attempt to run logistic regression to predict the strength of the effect of each individual's independent variable on the dependent variable, which is the odds ratio of vaccination intentions. We additionally use a classification methodology called random forest to figure out which variable(s) contribute most to the final model.

The ultimate goal is to develop a prediction model that is generalized and helpful to estimate the effect of variables in case of future pandemics like Covid-19. In order to test the

predictive power of the selected model estimates, we further compute the data regarding the decisions on the seasonal flu vaccine uptake within the same population by making certain modifications to the data to accommodate. How well the model fits the data can be checked through the ROC curve. If the curve yields higher sensitivity and specificity, we can conclude that the model we generated possesses generalization properties. If the logistic regression results in a higher AUC score, the estimates can effectively tell us the result of the relationship between the variables and their outcome. Eventually, the selected model will be further used to predict Covid-19 vaccination decisions. Moreover, the outcomes can guide future research on Covid-19 vaccination and public health efforts throughout the U.S.

1.3 Main Findings and Limitations

In general, the random forest model yields a higher AUC score and is utilized as the prediction model with some variable accommodations. The model suggests that the variables—attitude towards vaccines and state of residence—possess the most substantial explanatory power. The first variable is conventionally reasonable since the self-subjective consciousness has a solid and direct impact on decisions regarding vaccination after the roll-out. The second variable captures the variation in geographic elements. Accuracy statistics state that it largely contributes to the vaccination intentions. By further performing an analysis on the state of residence, we are able to measure the region-based effects, which play an essential role in developing specific policies and instructions in each state. However, there exist two unignorable limitations that may impact the prediction power of the model. The first limitation is driven by the strong positive correlation between the status of H1N1 vaccination and the seasonal flu shot in model generalization, which may generate instability in the model performance. The second one is the lack of continuous independent variables and hence the loss of information in model construction.

Section 2 describes the model specifics, variable selection, and necessary assumptions. It also introduces the mechanism of random forest from the machine learning models. Section 3 explains the data source and statistics. Section 4 compares the models through accuracy statistics and tests the prediction power of the selected model. It further discusses two variables that exert the most significant effect on the outcome. Section 5 concludes and outlines the weakness and next steps. All the tables and figures are included in Section 6.

2. Econometric Models

2.1 Binary Logistic Regression

2.1.1 Model Specifics

One model employed in this research is the logistic regression model, a typical form of binary response modeling. In specific, this type of model explores the relationship between a categorical response and a set of potentially influential predictor variables. Since this research aims to explore the association between individual characteristics and vaccination choices, our dependent variable is modified to indicate whether the respondent has received his/her vaccine, with 1 being vaccinated and 0 being unvaccinated. Due to this binary nature, a binary logit regression model will assist us in evaluating and measuring the effects of individual characteristics on people's vaccination willingness.

Accordingly, the proposed statistical model is as the following:

$$log\,\frac{p_i}{1-p_i} = logit(p_i) = \beta_0 + \beta_i * X_i', \text{ where}$$

- p_i is the probability that the respondent receiving the vaccination against the H1N1 virus
- β_i represent the coefficients corresponding to each covariate X_i; β_0 stands for the log odds ratio of the reference group
- X_i is the vector of covariates consisting of socio-demographic characteristics
- The RHS is the log odds of the probability of vaccination

2.1.2 Variable Selection

All the quantifiable responses of the survey that can be used as explanatory variables are included in the model as independent predictors. The model adopts ordinal and categorical variables by assigning numbers to each level, where each value denotes different meanings and is in increasing order. As for other nominal values that include multiple classes, such as state of residence, abbreviations are used to describe each class instead of numeric values since results of quantified but unordered indicative variables are not interpretable.

In specific, the models select from a pool of variables, including

- Personal attribute: age, gender, education level, marital and child status, past illness
- Economic status: family income, occupation, insurance coverage, work status
- Virus-related opinion: attitudes toward vaccines and the virus itself, knowledge of virus, doctor recommendation status
- Geographic factor: state of residence, indicator of metropolitan area

Utilizing the above mentioned predictors, the model will possess the ability to investigate how people in different groups react to the vaccination choices. With qualitative factors like states, one may explore the difference in log odds on vaccination probability between respondents from different states and the reference group. To enumerate, if Alabama is the reference state, with coefficients derived by the logit model, one can examine how California respondents may differ from those in terms of vaccination intention. In this paper, we treated the geographic factors as control variables and derived several specifications to analyze their importance since we cannot directly compute their effects as one value. By the same token, indicators and ordinal variables provide access to observe how the log odds respond regarding the movement among varying groups. The model will generally divide existing respondents into different groups and effectively measure how their socio-demographics shape vaccination intention. With new data, the model can also contribute to

providing credible predictions on probabilities. Therefore, logistic regression is an appropriate model for both effect interpretations and probability predictions.

Furthermore, this paper will implement variable selection measures to winnow the best variables for model construction. Typical measures including Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC)[1] will be applied in the process. Both measures are maximum likelihood estimate driven while adopting disparate penalties to combat overfitting problems. According to academic studies, some researchers claim that AIC and BIC are designed for different tasks. AIC aims to find the best-approximating model applicable to the unknown data generation process, while BIC tries to find the true model among candidate models. (Vrieze 2012) In other words, AIC is more appropriate for prediction while BIC performs better in explanation. Since the primary purpose of this paper is to build a predictive model, AIC will be the dominant selection criteria.

2.1.3 Model Assumptions

The binary logit model must fulfill the following assumptions to attain valid estimates.

1. Binary Response Variable

A proper binary logistic model requires a binary categorical dependent variable, where the number of outcomes must be two. Our model adopts an indicator variable to represent whether a person is vaccinated, which only has two unique outcomes values. The vaccination choice is 1 if the person has taken at least one dose of vaccine against H1N1 and 0 if never taken vaccination. Hence, the assumption is met.

2. Independence of Data

According to the raw data description, the survey drew respondents independently within each sample frame and recorded each once as one observation. Hence, the independence criteria are satisfied given the context.

[1] Both AIC and BIC are measures used to select models. The smaller, the better.

3. No Influential Observations (Outliers)

The model assumes no extreme observations in the data set. Hence, the model diagnostic procedure will employ measures like Cook's distance to measure the influence of each data point and remove the influential outliers to ensure no distortion in the model accuracy. The corresponding process is exhibited in the Appendix 6.7.

4. Linearity Assumption

In general, logit models assume a linear relationship between the logit of the dependent variable and continuous explanatory predictors, which can be tested through the Box-Tidwell test. However, the original survey data collected all the information as categorical values, in binary, nominal and ordinal. As logit regression does not require continuous variables, no further steps are required here.

5. No Multicollinearity

Logit models assume the absence of severe multicollinearity among the independent variables, implying no high correlations between two or more predictors. Strong multicollinearity will reduce model accuracy by undermining the statistical significance of explanatory variables. Variance Inflation Factor (VIF) [2] is an approach to measuring multicollinearity in the proposed independent variables. The corresponding diagnostic process is exhibited in the Appendix 6.7.

2.2 Machine Learning: Random Forest

2.2.1 Introduction to Random Forest

Considering that most explanatory variables and the responding variables are categorical, this paper will examine if a classification method is more suitable than regression models because it does not require any strong assumption on the distribution of the residuals or the linearity of the true model. Random forest is a commonly used classification

[2] Typical cutoffs for VIF are 5 and 10. Variables with VIF under the cutoff are proved to have no strong multicollinearity.

methodology in machine learning. It is a collective classification model of various decision trees. "Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features(Scikit Learn, 2007-2022)." The random forest algorithm generates from 1 tree to a maximum number of n trees where n can be chosen. The model will generate each decision tree using a random subset of auxiliary variables. For each number of trees from 1 to n, a collective forest of the trees is generated and used to predict the response variable. In most cases, the training set (as is discussed in section 3.3) error rate will decrease at a decreasing rate as the number of trees increases. The algorithm will also select a different number of explanatory variables for each decision tree and come to the optimal number itself.

The fundamental mechanism of this algorithm is straightforward since Python and R can run the function and do all the math backstage. It uses the training dataset to generate an optimal model with minimum error in predicting the training dataset. The testing data set will then be used to test the model's accuracy and help with model selection. The randomness of the training and testing dataset ensures the representativeness of the model. In other words, randomness makes sure that the out-of-bag variance will not be driven by sample selection bias ("bag" refers to the training dataset, "out-of-bag" observations are observations outside of the training dataset). Holding everything else equal, the variations in the vaccination decision in the training dataset can be almost fully captured by the model and explained by the explanatory variables included in the model when predicting a new data set.

However, there are some concerns in the empirical analysis. First of all, the results from the random forest are not interpretable. The model is insufficient to trace what explanatory variables specifically caused the variations in the responding variables. Neither can it generate interpretable correlation coefficients. Secondly, to ensure a certain level of prediction accuracy,

the sample that's being predicted must be similar to the training set in terms of macro factors. For example, the training set only includes observations from the U.S. If the model trained by this training set is used to make predictions on samples from Asian countries that have very different macroeconomic factors, the results are likely to be inaccurate because people in certain Asian countries might make decisions based on factors that are very different from the explanatory variables included in the model. This statement can be generalized as the insufficiency in random forest imputation caused by highly skewed and/or missing data.

2.2.2 Model Specifics

The random forest algorithm inputs must include explanatory variables and the number of trees upon choosing. The maximum number of trees does not affect the outcome significantly because the error rates of both training and testing sets will remain stable after a certain number of trees. As long as the maximum number of trees exceeds that threshold, the algorithm will be able to capture the optimal number of trees by comparing all the forests.

All the explanatory variables sorted for the logit model are included in the random forest model. The more a variable is used as nodes or splitting points, the more critical that variable is. In other words, the model is self-selected with all the input variables. Even though the model results are not highly interpretable, one way to assess the significance of each variable is to generate a mean decrease Gini coefficient for each variable from the model. This coefficient measures the extent to which each variable contributes to the homogeneity of each node. Permuting a relatively useful variable for classification would decrease the Gini gain significantly. The higher the mean decrease Gini of a variable is, the more influential the variable is in the forest. These coefficients are highly interpretable and have strong economic significance. They also provide useful guidance for treatment effect analysis because they indicate which variables are important and worth further analysis. However, while the random

forest model might be a sound methodology for prediction, the logit model would be a better choice to control for the average treatment effect of a specific variable.

Several predictors randomly sampled at each split for an individual tree are set to default in R, which is the square root of the total number of predictors. Different numbers of trees will be tested to identify the optimal number of trees (i.e., 1000, 500, 100). The dependent variable is the H1N1 vaccination status which takes 0 or 1, indicating whether or not vaccinated.

The performance of the model will be evaluated through the area under the curve (AUC[3]) using the testing data set (section 3.3). AUC will be used to compare the in-sample accuracy of the random forest model with that of the logit model. Furthermore, the model trained for H1N1 vaccination will be used to predict the vaccination status of seasonal flu by inputting the same explanatory variables surveyed for seasonal flu. The validity and limitation of this part of the analysis will be discussed later, while it provides some insights if these socio-demographic characteristics can be generalized to predict vaccination decisions for other comparable vaccines.

3. Data

3.1 Raw Data

Our research employs a subset of the National 2009 H1N1 Flu Survey (NHFS), which is a national-wide telephone survey conducted by the Centers for Disease Control and Prevention (CDC) in late 2009 and early 2010. The survey targeted all residents that are at least six months old and live in the United States at the time of the interview. It applied the geographically stratified sampling method, where respondents are randomly selected from each eligible household within each stratum. The data provided access to diverse information, including respondents' vaccination status, knowledge, attitudes, and practice toward

[3] AUC (Area Under the Curve) is the total area under the curve ROC (Receiver Operating Characteristics). AUC-ROC curve provides measurements for classification model performance at various threshold settings. It quantifies the extent to which the model can differentiate different classes. The higher AUC is, the better the model can classify each observation as its true or observed class.

pandemics and vaccines, past respiratory illness, and detailed socio-demographics. Table 1 exhibits the mean differences of the crucial characteristics between the vaccinated and unvaccinated groups to explore characteristics associated with vaccination acceptance. Similar data is also collected for seasonal flu vaccination. Detailed descriptions of all variables used are included in Appendix 6.1.

3.2 Summary Statistics

As shown in Appendix 6.1: Table 1, there are 23515 observations with around 30% H1N1 vaccination rate. On average, the data suggested that non-vaccinated respondents possess consistent and similar characteristics such as lower family income, lower education level, younger age, and fewer concerns about the virus. Likewise, the negative and significant differences in the indicator variables imply that health workers, insurance-covered people, and married couples have a higher willingness to take vaccinations (Table 1). In terms of subjective perspectives such as attitudes, the comparatively large and significant differences signify the discrepancies in people's awareness and knowledge of both the virus and the vaccine. Note that the differences in the child indicator are statistically insignificant, suggesting similar vaccination likelihood for people with and people without children. Further specifications will be derived from speculations of potential state-level or regional disparities by examining their potential existence.

3.3 Train-test Split Technique

This study also aims to construct a model that can relatively precisely predict people's vaccination intention through the past H1N1 vaccine acceptance pattern. Hence, the model will apply the train-test split technique for performance evaluation to refine the best prediction model. To extend, data sets were randomly divided into two subsets, referred to as training and testing sets. The training dataset is utilized to generate models, while the testing data set is created for comparison. Specifically, models generated from the training set will be used to predict the class of the testing set. The prediction accuracy is reflected via AUC, as mentioned

in section 2.2.2. In addition, the randomness in splitting the testing and training sets ensures symmetry in terms of omitted characteristics that are not included in the model, which substantially validates AUC estimates as measurements of accuracy. This method can effectively reduce the potential existence of the overfitting problem, enhancing the credibility of model deployment in future events.

4. Results

4.1Main Results

After running the econometric models mentioned above using R, the results from the logistic model and random forest are not significantly different. Comparing the resulting AUC from AIC, BIC, and full model, the final model for logistic regression is shown in Table 3 in section 6.3 as "Full". The random forest model results that include all the explanatory variables are shown in Table 2 and the following Figure 1 and 2. The AUC estimates of the two models are very close to each other (Figure 2: Random Forest AUC = 0.786; Figure 6: Logistic AUC = 0.798). The logistic model has a higher AUC, thus more accuracy when predicting the test set for H1N1. As random forest results are not interpretable and show lower AUC, the logistic regression will be chosen to identify the significance of the explanatory variables and their correlations with the H1N1 vaccination decision.

Table 2 shows the explanatory power of each variable from the random forest model. Both models suggest that state of residence is one of the most important variables and significantly influences the resulting class. This indicates that the vaccine roll-out plan is not consistent across the country regarding time and availability. This specific variable will be further discussed in section 4.3. In addition, the most significant variable from both models is the attitude towards vaccines. The more acceptance people have towards vaccines, the more likely they will get vaccinated. Income also showed significance in the logistic model and is positively correlated with vaccination decisions (Table 3). By controlling for attitude and

income variables, the result suggests that to get a higher vaccination coverage, the U.S. government needs to enhance vaccination education for people with lower income, which is viable through various median intermediates such as social media platforms and TV streaming companies.

4.2 Generalization

In order to compare the generalization power of the models, both models were used to predict the vaccination decision for seasonal flu shots. However, the data does not include the knowledge of vaccines or the attitude towards viruses for seasonal flu. Thus, the two models were reconstructed by deleting these two exogenous variables using the same training data set for H1N1. In order to ensure the same level of accuracy of the models before applying them to seasonal flu vaccination data, AUC estimates of the newly constructed models were generated from the testing set of H1N1 (Figure 7: Logistic AUC = 0.791; Figure 4: Random Forest AUC = 0.782). These estimates are still very close to the estimates of the original models that contain the two deleted variables.

Using both new models to predict the seasonal flu shot decision, the random forest model performed better than the logistic model (Figure 8: Logistic AUC = 0.815; Figure 5: Random Forest AUC = 0.826). The random forest model not only performed better than the logit model in this case but also scored a higher AUC than it initially did when tested on the testing set for H1N1 data. This could potentially be attributed to overfitting as two exogenous variables were deleted from both models. However, AUC estimates of the new models are lower than the models before deleting the two variables, which would be the opposite case had the better performance in seasonal flu data been caused by overfitting. Another possibility would be that there exists a strong correlation between the seasonal flu shot decision and the H1N1 vaccination decision, which is discussed in section 5.2.

4.3 State of Residence

Attitude being the most significant variable conventionally makes sense since people with less negative opinions towards vaccines will get vaccinated faster after the rollout. This part aims to discover why the state of residence would be an important, influential factor in vaccination decisions. For better illustration, we delineated the H1N1 vaccination rates of each state on the U.S. map. Referring to Figure 9, the vaccination coverage is exhibited through changes in the shades of color, where higher vaccinated states possess darker orange, and low vaccinated areas use lighter orange. It is clearly demonstrated that clusters of high vaccination coverage areas and low-coverage regions exist. A possible explanation is that adjacent states are prone to have comparable fundamentals such as population composition in terms of age, race, income, culture, religion, and geographic conditions, indicating regionalism. For instance, states in the southern region, such as those around Mississippi, tend to have low popularization of vaccination. As these states are acknowledged to suffer from poverty, one may suspect low accessibility to health-care services in those regions or high commuting costs of arriving at the vaccination location. This is empirically testable if sufficient vaccine registration data from different locations are available. Simultaneously, since the clusters display fading ripples, another reason may be derived from the vaccine rollout. In other words, the timing of introduction and implementation of vaccination could differ from state to state and thus affect the availability to residents.

To validate the effect of the geographical element, we computed four logistic specifications by applying two different sets of control variables, state of residence and MSA (metropolitan statistical area). According to Table 3, both MSA and state largely contribute to the elements affecting vaccination intention in the presence of lower AIC values. Adding these factors, the increased significance in some variables further validated the model improvements. However, the decrease in AIC is larger when controlling for the state, and there is no significant

decrease in AIC by adding MSA when the model is already controlling for the state. MSA is still a significant factor if the analysis were focused on exploring the vaccination coverage differences between cities and suburbs.

Further exploration of the state effect corroborates our preliminary speculations. For instance, the state coefficient for Mississippi is statistically significant and negative. To interpret, holding all else equal, the log odds ratio of vaccination probability in Mississippi residents is 1 unit lower than those from the reference state of Arkansas. Specifically, the model omitted certain factors causing the state differences and disparities.

As an extension to the economic implications above, the states with the highest (R.I.) and lowest (M.S.) vaccination rates were selected. H1N1 vaccination decision conditioned on these two states is regressed respectively, as exhibited in Table 4. Excluding the attitude factor mentioned in the previous section, the results from the two regression models showed that the significance level of the same predictor varies across states. For instance, medical advice on the vaccine is a significant, influential factor in Mississippi, while age and employment status in health facilities are the primary factors in Rhode Island. Accordingly, if the Mississippi government wants to enhance the spread of vaccination, it should focus on campaigns that incentivize doctors to educate about related detriments of the virus and the benefits of getting the vaccination.

5. Conclusion

5.1 Main Conclusion

In summary, this paper utilizes logistic regression models and random forest algorithms to construct a prediction model on people's vaccination intention, given their socio-demographics, including individual attributes, economic situation, opinions, and geographic factors. Through various specification selection processes, our random forest model achieved an AUC score of 0.826 when generalized to predict other pandemic vaccination statuses. In

addition, we performed a detailed analysis on one of the influential factors, state of residence, using the logit model as it has better in-sample performance. The results suggest that implementing region-based policies targeting low-vaccination rate areas is more efficient because people in different states make vaccination decisions based on different factors. Besides, to enhance the overall vaccination willingness in the U.S., the government authorities should place more emphasis on people's attitudes towards the vaccines.

5.2 Caveat & Further Steps

One notifiable limitation of our research comes from the process of generalization. In this paper, we established the externality of our model by extending it to another set of data. Although the model attained an AUC score of 82%, there exists a strong positive correlation (Spearman test result: 0.4462) between the vaccination indicators of H1N1 and seasonal flu. The high AUC score might be attributed to this strong correlation. There is a potential that the model will perform poorly when predicting COVID-19 data if the covid-19 vaccination status is not strongly correlated with H1N1 vaccination status. However, this is empirically testable because the random forest model can always be reconstructed by adding or deleting factors without needing to make any distributional assumptions. To mitigate the problem and further extend the analysis, new data from other pandemics should be acquired. If COVID-19 vaccination predictions for the same group of people surveyed for this paper can achieve a consistently high rate, a more perfected model can be generalized for improving vaccine coverage of future pandemics.

Another limitation of the logistic regression model is that due to the raw data response setting, we only have categorized variables to employ as independent variables. As discussed in *Regression modeling strategies* (Frank Harrell, 2015), logit regression would usually prefer continuous variables to categorized numerical variables since they would contain more information contributing to model construction. For instance, our logit model performance

could be significantly enhanced by numerical values for ages instead of the age group indicators as long as we satisfy the linearity assumption between the variable and the log odds.

Lastly, the model included in this paper did not consider the time stamp of vaccination. It only measures whether an individual will get vaccinated or not after one year of the roll-out plan, but not how fast this particular person will get vaccinated. Longitude data analysis can measure the time differences in getting vaccinated.

6. Appendix

6.1 Data Summary

Table 1: Summary Statistics by H1N1 Vaccination Status

	(1) Vaccinated		(2) Nonvaccinated		(3) Difference
	mean	sd	mean	sd	
Inc	5.231	1.911	4.944	1.965	-0.287***
Edu	3.150	0.953	2.973	0.990	-0.178***
Health_Work	0.211	0.408	0.082	0.274	-0.129***
Insure	0.940	0.237	0.858	0.349	-0.083***
Age_grp	3.251	1.425	3.014	1.437	-0.237***
Sex	0.401	0.490	0.419	0.493	0.018**
Attitude_h	2.794	0.858	2.416	0.862	-0.377***
Marriage	0.577	0.494	0.515	0.500	-0.062***
Attitude_vacc_h	2.574	0.906	1.757	0.769	-0.816***
Knowledge_h	2.425	0.580	2.209	0.597	-0.215***
Rec_h	0.067	0.249	0.024	0.152	-0.043***
Ill	0.363	0.481	0.259	0.438	-0.104***
Child	0.543	0.924	0.553	0.942	0.010
Observations	7067		16448		23515

Notes: *Inc*: 1:$10k, 2:10k-$15k, 3:$15k-$25k, 4:$25k-$35k, 5:$35k-$50k, 6:$50k-$75k, 7:$75k-$100k; *Edu*: 1:< 12 years, 2: 12 yrs, 3: some college, 4: College Graduate; *Health_Work*: whether the respondent is a health worker; *Insure*: whether the respondent is insured; *Age_grp*: 1:18 - 34 yrs, 2:35 - 44 yrs, 3: 45 - 54 yrs, 4:55 - 64 yrs, 5:65+ yrs; *Sex*: whether the respondent is male; *Attitude_h*: 1:not concern about H1N1 flu at all, 2:not very concern, 3:somewhat concern, 4:very concern; *Marriage*: whether the respondent is married; *Attitude_vacc_h*: 1:very low necessity of getting H1N1 vaccination, 2:somewhat low necessity, 3:somewhat high necessity, 4:very high necessity; *Knowledge_h*: 1:no knowledge about H1N1 Flu, 2:a little knowledge, 3:know very well; *Rec_h*: whether the respondent receives recommendation from doctor; *Ill*: whether the respondent has chronic medical condition; *Child*: whether the respondent has child
* p<0.05 ** p<0.01 *** p<0.001

6.2 Random Forest Main Results with full explanatory variables

Table 2: RF: Variable Importance

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
Inc	11.769	6.358	14.261	505.389
Edu	2.647	8.131	7.575	344.526
Health_Work	10.544	21.181	23.495	178.581
Insure	8.587	12.225	14.040	104.271
Age_grp	12.278	9.680	16.668	408.611
Sex	3.381	0.474	3.383	180.624
Attitude_h	21.683	-2.017	17.327	364.615
Attitude_vacc_h	78.059	80.053	104.772	1,203.917
Knowledge_h	1.229	18.146	14.463	253.142
Marriage	3.467	4.000	5.597	171.170
State	-0.253	0.316	0.011	1,090.879
Rec_h	21.350	17.774	26.942	114.495
Ill	3.417	6.264	6.612	151.735
Child	3.140	10.488	8.188	153.285
Employ	2.453	6.175	6.267	202.646
Metro	2.736	1.571	3.230	328.495

Figure 1.

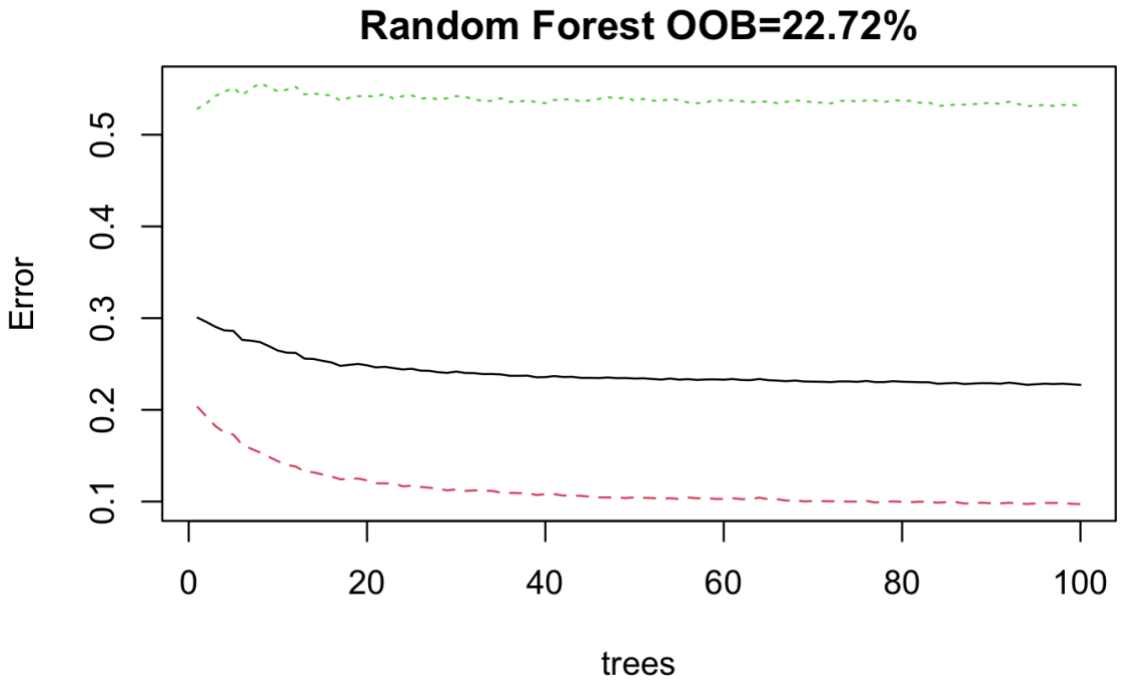
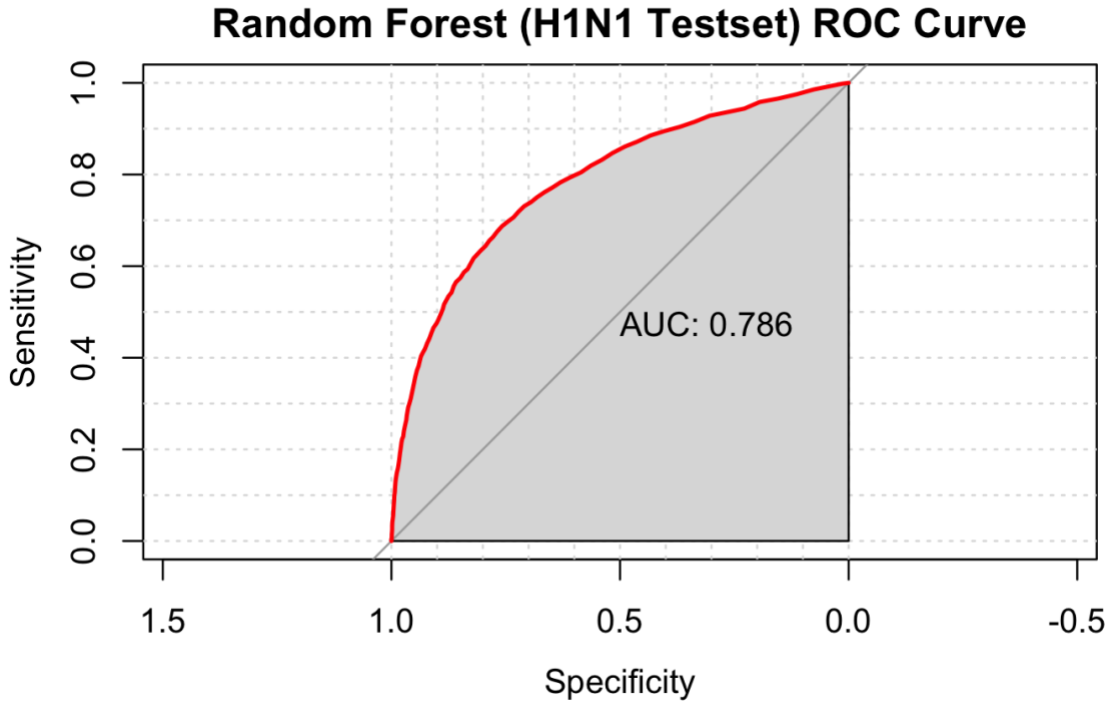


Figure 2.



6.3 Random Forest Main Results with full explanatory variables

Figure 3.

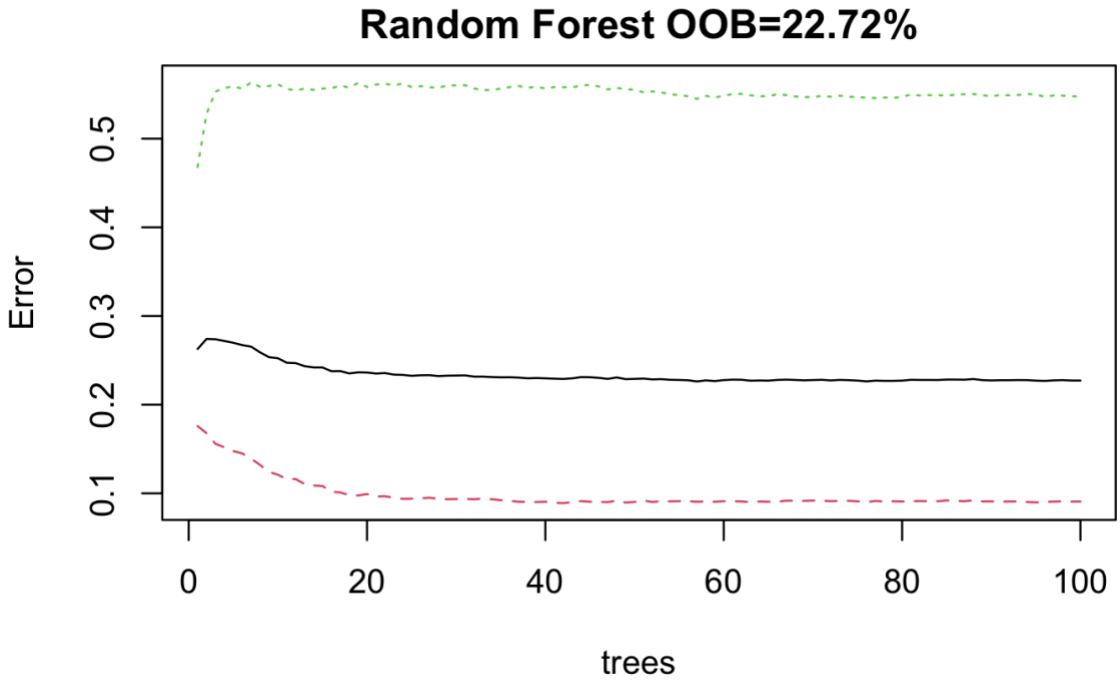


Figure 4.

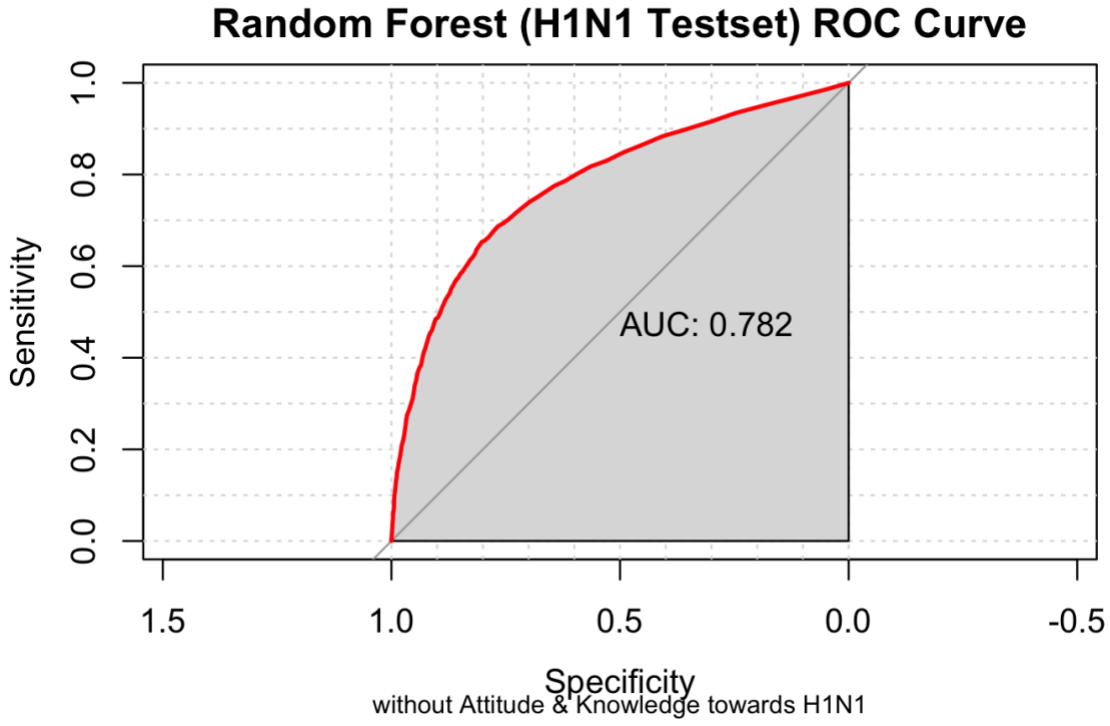
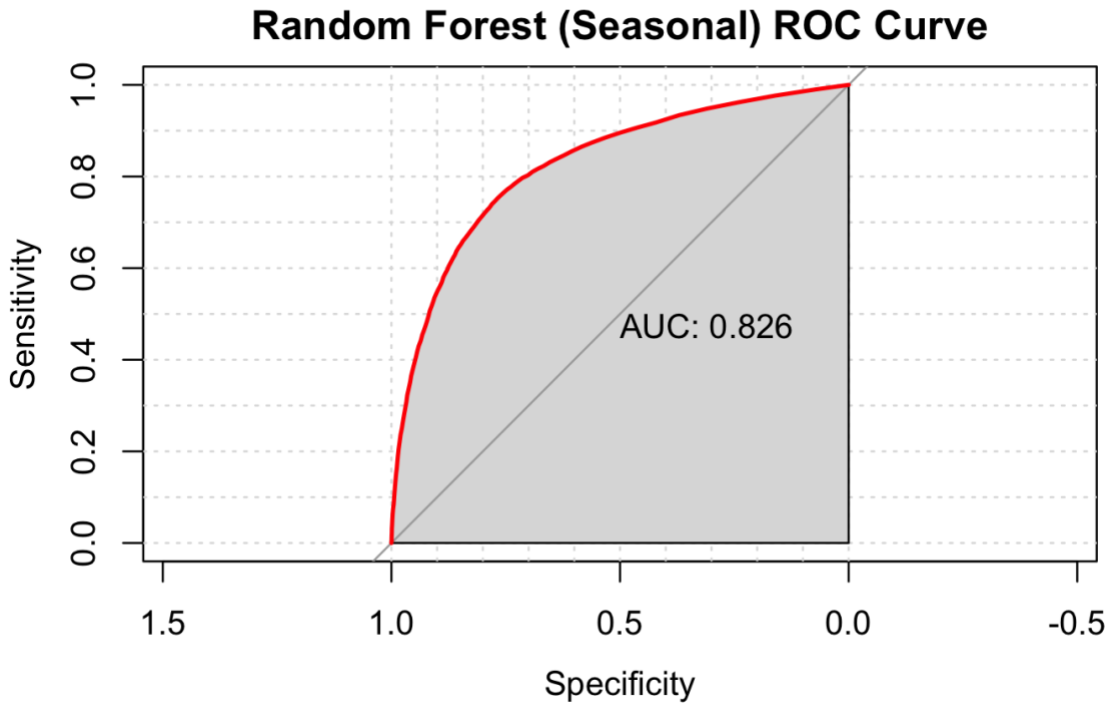


Figure 5.


Random Forest (Seasonal) ROC Curve
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
AUC: 0.826
1.5
1.0
0.5
0.0
-0.5
Specificity

6.4 Logit Regression Main Results with full explanatory variables

Table 3: Model Results

	Dependent variable:					
	Vacc_h					
	None	Control: MSA	Control: State	BIC	AIC	Full
	(1)	(2)	(3)	(4)	(5)	(6)
Inc	0.029**	0.031**	0.028**	0.040***	0.029**	0.029**
	(0.014)	(0.014)	(0.014)	(0.013)	(0.014)	(0.014)
Edu	0.170***	0.168***	0.169***	0.165***	0.167***	0.167***
	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
Age_grp	0.158***	0.159***	0.156***	0.149***	0.157***	0.157***
	(0.018)	(0.018)	(0.018)	(0.016)	(0.018)	(0.018)
state_mapAL			-0.383*		-0.370*	-0.370*
			(0.203)		(0.203)	(0.203)
state_mapAR			0.073		0.085	0.085
			(0.199)		(0.199)	(0.199)
state_mapAZ			-0.250		-0.249	-0.249
			(0.193)		(0.193)	(0.193)
state_mapCA			0.103		0.104	0.104
			(0.190)		(0.191)	(0.191)
state_mapCO			-0.290		-0.289	-0.289
			(0.202)		(0.203)	(0.203)
state_mapCT			-0.247		-0.220	-0.220
			(0.203)		(0.203)	(0.203)
state_mapDC			-0.033		-0.089	-0.089
			(0.187)		(0.189)	(0.189)
state_mapDE			-0.046		-0.015	-0.015
			(0.195)		(0.195)	(0.195)
state_mapFL			-0.177		-0.162	-0.162
			(0.202)		(0.203)	(0.203)
state_mapGA			-0.397**		-0.368*	-0.368*
			(0.201)		(0.202)	(0.202)
state_mapHI			0.413**		0.427**	0.427**
			(0.191)		(0.191)	(0.191)
state_mapIA			0.286		0.298	0.298
			(0.191)		(0.192)	(0.192)
state_mapID			-0.076		-0.068	-0.068
			(0.204)		(0.205)	(0.205)
state_mapIL			-0.325		-0.308	-0.308

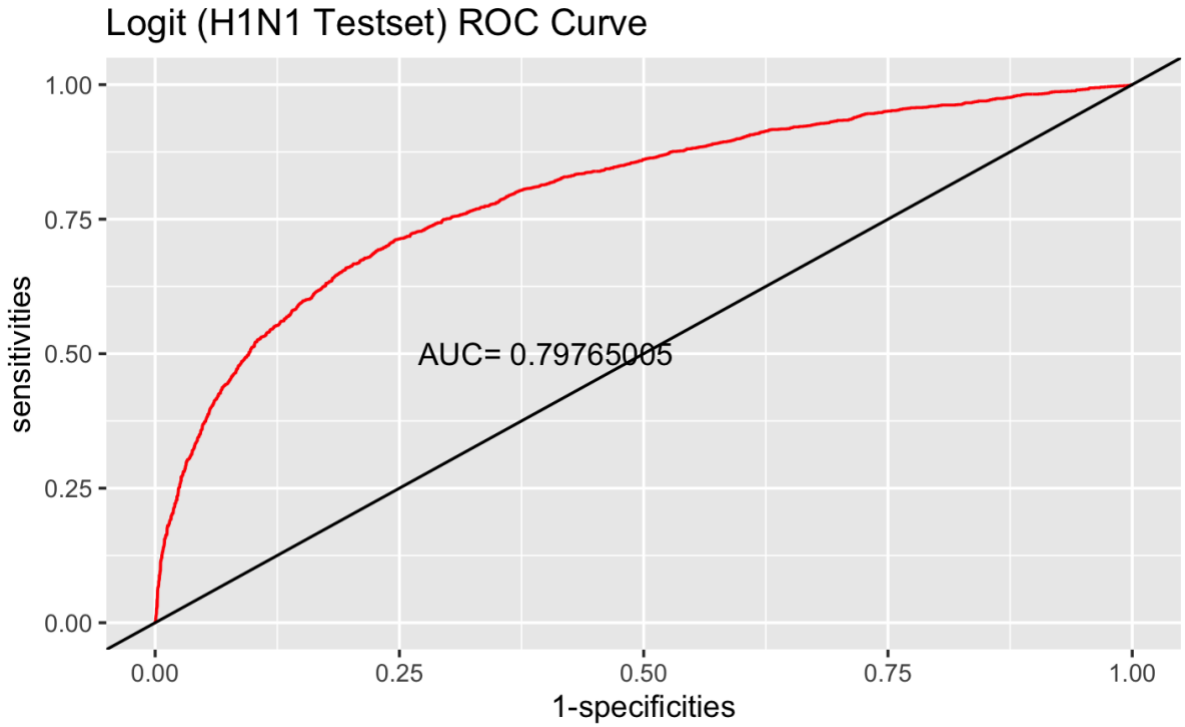
	(0.201)	(0.201)	(0.201)
state_mapIN	-0.214	-0.205	-0.205
	(0.200)	(0.200)	(0.200)
state_mapKS	-0.0002	0.012	0.012
	(0.195)	(0.195)	(0.195)
state_mapKY	-0.220	-0.195	-0.195
	(0.207)	(0.207)	(0.207)
state_mapLA	-0.554***	-0.538***	-0.538***
	(0.209)	(0.209)	(0.209)
state_mapMA	0.575***	0.600***	0.600***
	(0.194)	(0.195)	(0.195)
state_mapMD	-0.125	-0.100	-0.100
	(0.197)	(0.198)	(0.198)
state_mapME	0.220	0.244	0.244
	(0.193)	(0.193)	(0.193)
state_mapMI	-0.353*	-0.331	-0.331
	(0.209)	(0.209)	(0.209)
state_mapMN	0.293	0.311	0.311
	(0.193)	(0.193)	(0.193)
state_mapMO	-0.235	-0.217	-0.217
	(0.209)	(0.210)	(0.210)
state_mapMS	-0.926***	-0.905***	-0.905***
	(0.219)	(0.219)	(0.219)
state_mapMT	-0.182	-0.171	-0.171
	(0.208)	(0.209)	(0.209)
state_mapNC	-0.110	-0.101	-0.101
	(0.195)	(0.195)	(0.195)
state_mapND	0.260	0.264	0.264
	(0.199)	(0.199)	(0.199)
state_mapNE	0.166	0.170	0.170
	(0.195)	(0.195)	(0.195)
state_mapNH	0.161	0.183	0.183
	(0.204)	(0.205)	(0.205)
state_mapNJ	-0.764***	-0.721***	-0.721***
	(0.205)	(0.207)	(0.207)
state_mapNM	0.061	0.057	0.057
	(0.186)	(0.186)	(0.186)
state_mapNV	-0.208	-0.204	-0.204
	(0.199)	(0.200)	(0.200)
state_mapNY	-0.561***	-0.558***	-0.558***
	(0.203)	(0.203)	(0.203)

state_mapOH			-0.188		-0.161	-0.161
			(0.205)		(0.206)	(0.206)
state_mapOK			-0.214		-0.203	-0.203
			(0.202)		(0.202)	(0.202)
state_mapOR			0.021		0.025	0.025
			(0.198)		(0.198)	(0.198)
state_mapPA			-0.320		-0.294	-0.294
			(0.204)		(0.205)	(0.205)
state_mapRI			0.678***		0.702***	0.702***
			(0.191)		(0.193)	(0.193)
state_mapSC			-0.359*		-0.331	-0.331
			(0.203)		(0.203)	(0.203)
state_mapSD			0.562***		0.571***	0.571***
			(0.198)		(0.198)	(0.198)
state_mapTN			-0.412**		-0.409*	-0.409*
			(0.209)		(0.209)	(0.209)
state_mapTX			-0.025		-0.024	-0.024
			(0.189)		(0.189)	(0.189)
state_mapUT			0.187		0.209	0.209
			(0.193)		(0.193)	(0.193)
state_mapVA			0.172		0.184	0.184
			(0.187)		(0.187)	(0.187)
state_mapVT			0.246		0.273	0.273
			(0.197)		(0.198)	(0.198)
state_mapWA			0.136		0.153	0.153
			(0.198)		(0.199)	(0.199)
state_mapWI			0.046		0.067	0.067
			(0.195)		(0.195)	(0.195)
state_mapWV			-0.388*		-0.365*	-0.365*
			(0.209)		(0.209)	(0.209)
state_mapWY			0.084		0.096	0.096
			(0.198)		(0.199)	(0.199)
Health_Work	0.891***	0.891***	0.934***	0.885***	0.933***	0.933***
	(0.059)	(0.059)	(0.060)	(0.059)	(0.060)	(0.060)
Insure	0.808***	0.810***	0.789***	0.808***	0.789***	0.789***
	(0.080)	(0.080)	(0.081)	(0.080)	(0.081)	(0.081)
Sex	0.340***	0.338***	0.338***	0.337***	0.336***	0.336***
	(0.042)	(0.042)	(0.042)	(0.041)	(0.042)	(0.042)
Attitude_vacc_h	1.092***	1.092***	1.097***	1.093***	1.097***	1.097***
	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)
Attitude_h	0.143***	0.142***	0.154***	0.148***	0.153***	0.153***

	(0.025)	(0.025)	(0.026)	(0.025)	(0.026)	(0.026)
Knowledge_h	0.399***	0.401***	0.377***	0.401***	0.378***	0.378***
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
Marriage	0.075*	0.081*	0.072		0.079*	0.079*
	(0.045)	(0.045)	(0.046)		(0.046)	(0.046)
Child	0.066	0.071	0.074		0.078	0.078
	(0.051)	(0.051)	(0.052)		(0.052)	(0.052)
Ill	0.257***	0.259***	0.278***	0.252***	0.278***	0.278***
	(0.044)	(0.044)	(0.044)	(0.044)	(0.044)	(0.044)
Rec_h	1.223***	1.223***	1.217***	1.223***	1.215***	1.215***
	(0.096)	(0.096)	(0.098)	(0.096)	(0.098)	(0.098)
employ_grpnot in labor force	0.249***	0.252***	0.288***	0.250***	0.289***	0.289***
	(0.048)	(0.048)	(0.049)	(0.048)	(0.049)	(0.049)
employ_grpunemployed	0.129	0.132	0.155	0.130	0.154	0.154
	(0.093)	(0.093)	(0.094)	(0.093)	(0.094)	(0.094)
MSA_grpMSA, principle city		0.130***			0.110**	0.110**
		(0.047)			(0.050)	(0.050)
MSA_grpnon-MSA		0.075			0.024	0.024
		(0.049)			(0.054)	(0.054)
Constant	-6.937***	-7.015***	-6.863***	-6.907***	-6.923***	-6.923***
	(0.158)	(0.162)	(0.209)	(0.154)	(0.213)	(0.213)
Observations	16,461	16,461	16,461	16,461	16,461	16,461
Log Likelihood	-7,947.291	-7,943.326	-7,819.977	-7,950.173	-7,817.464	-7,817.464
Akaike Inf. Crit.	15,926.580	15,922.650	15,771.950	15,928.340	15,770.930	15,770.930

Note: *p<0.1; **p<0.05; ***p<0.01

Figure 6.


Logit (H1N1 Testset) ROC Curve
1.00
0.75
0.50
0.25
0.00
sensitivities
AUC= 0.79765005
0.00
0.25
0.50
0.75
1.00
1-specificities

6.5 Logit Regression Main Results omitting Knowledge_h & Attitude_h

Figure 7.

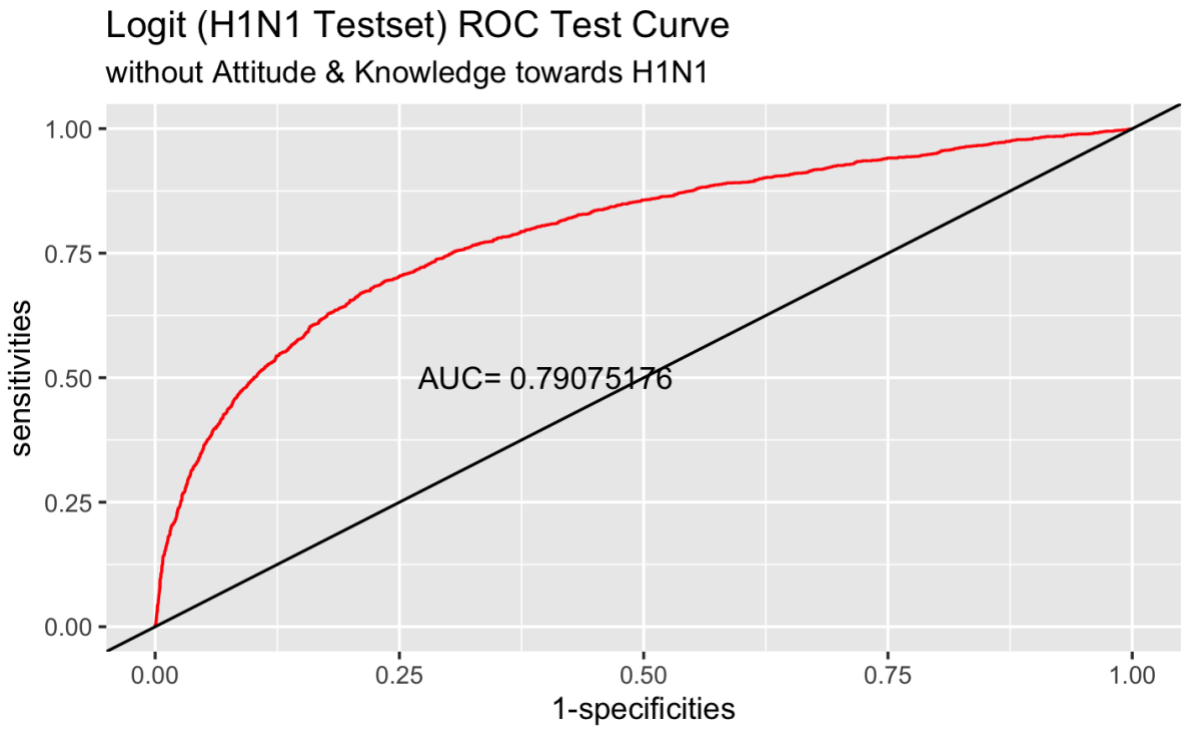
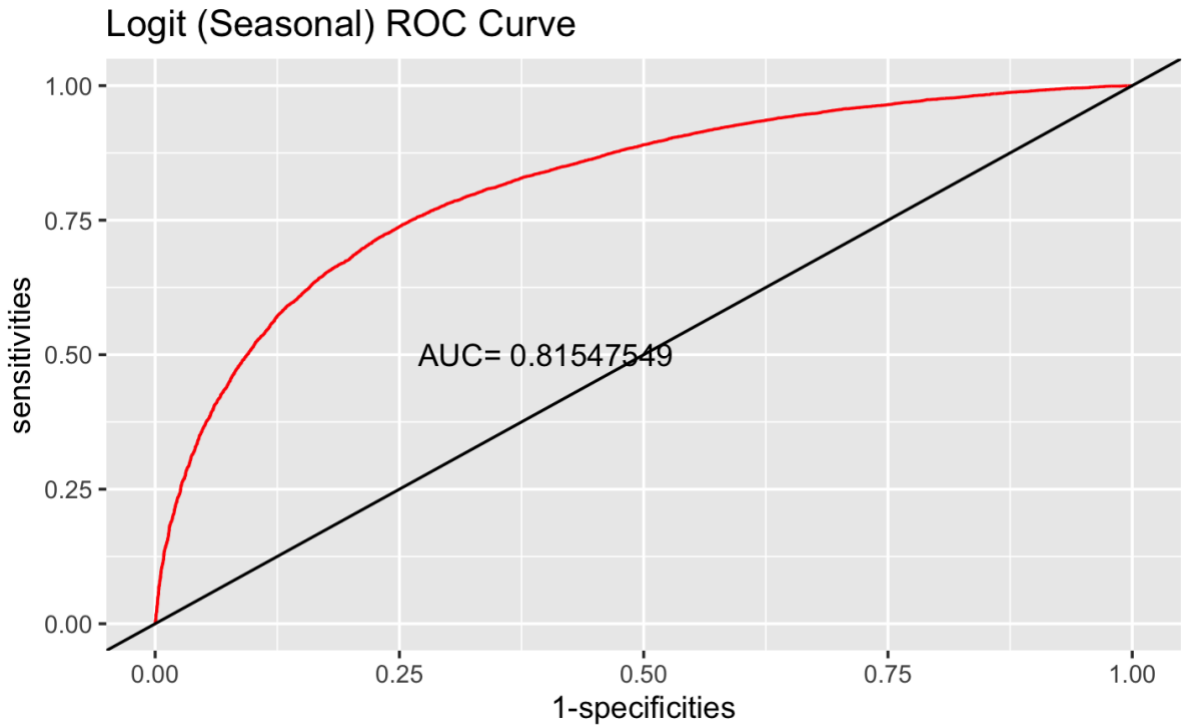


Figure 8.



6.6 State Analysis

Figure 9.

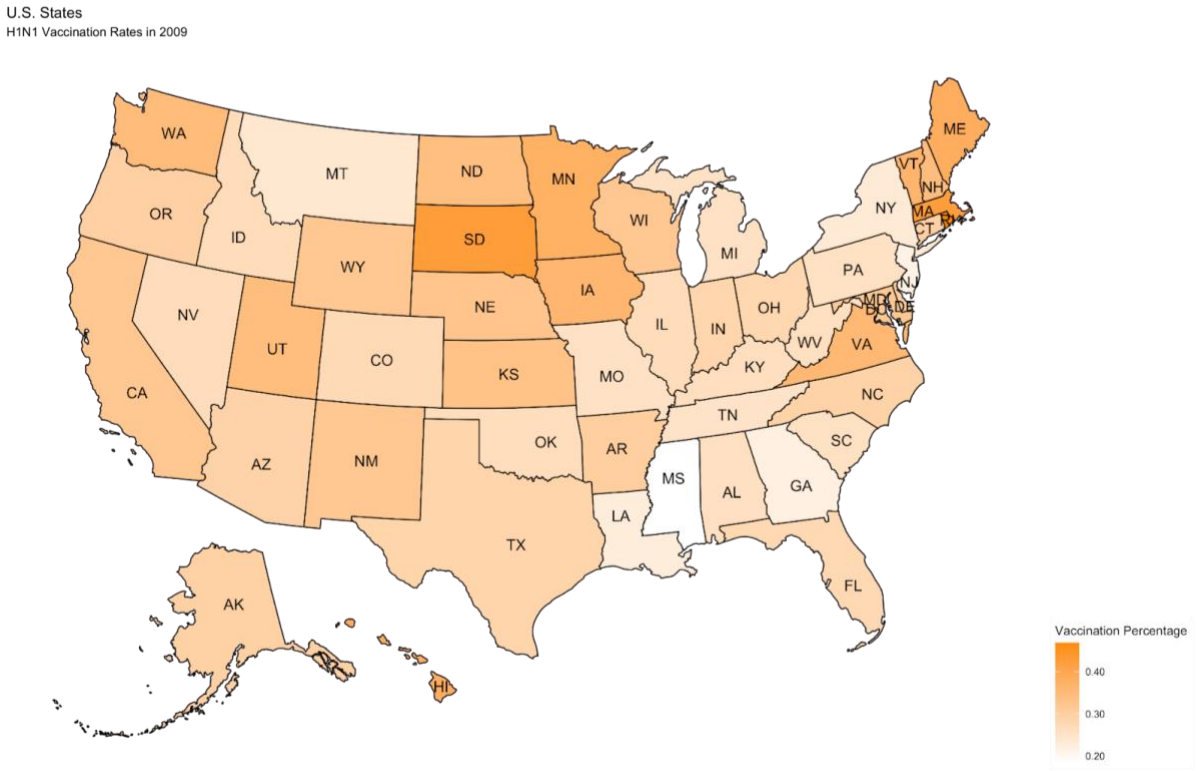


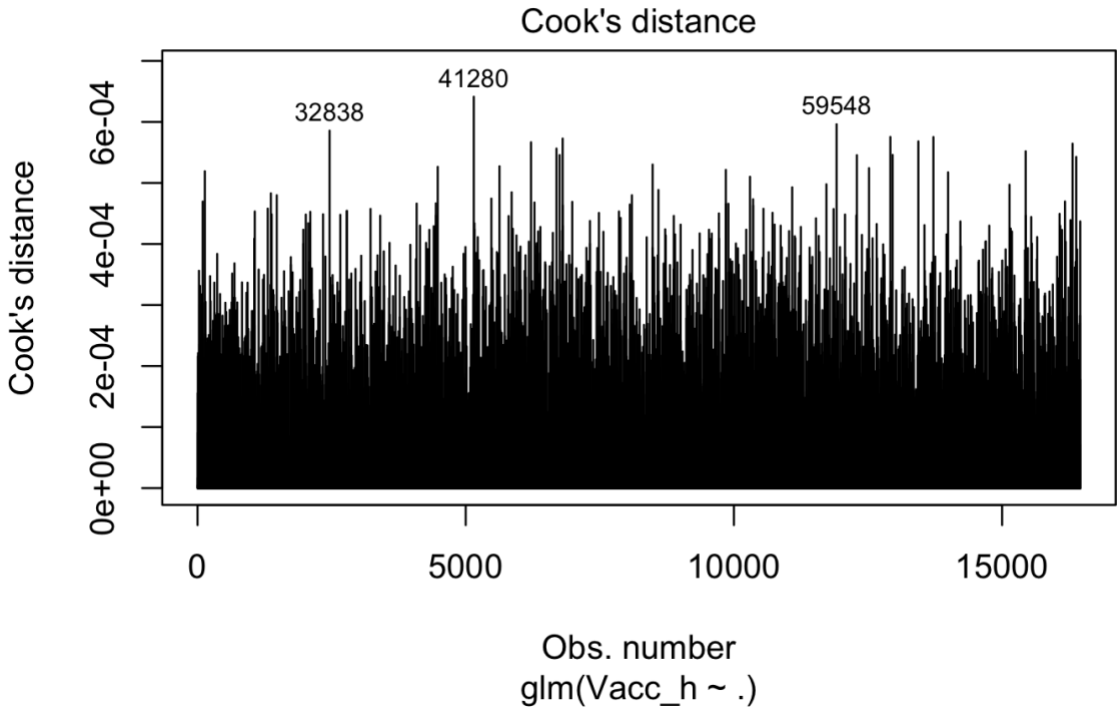
Table 4: Model Results

	Dependent variable:	
	Vacc_h	
	Mississippi	Rhode Island
	(1)	(2)
Inc	-0.107	0.080
	(0.101)	(0.086)
Edu	0.208	0.084
	(0.173)	(0.142)
Age_grp	-0.003	0.313***
	(0.122)	(0.107)
Health_Work	0.208	1.762***
	(0.433)	(0.411)
Insure	0.747	1.381**
	(0.488)	(0.555)
Sex	0.663**	0.391
	(0.296)	(0.250)
Attitude_vacc_h	0.790***	1.228***
	(0.156)	(0.170)
Attitude_h	-0.201	0.181
	(0.170)	(0.154)
Knowledge_h	0.371	0.087
	(0.249)	(0.215)
Marriage	0.017	-0.299
	(0.351)	(0.293)
Child	0.208	0.155
	(0.356)	(0.317)
Ill	0.240	0.250
	(0.298)	(0.267)
Rec_h	1.979***	0.262
	(0.603)	(0.494)
employ_grpnot in labor force	0.427	0.152
	(0.334)	(0.303)
employ_grpunemployed	-0.614	0.253
	(0.751)	(0.538)
MSA_grpMSA, principle city	-0.046	0.302
	(0.477)	(0.271)
MSA_grpnon-MSA	0.514	
	(0.323)	
Constant	-5.277***	-6.933***

	(1.069)	(1.020)
Observations	444	414
Log Likelihood	-176.986	-216.574
Akaike Inf. Crit.	389.973	467.148
Note:		*p<0.1; **p<0.05; ***p<0.01

6.7 Model Diagnostics (Logit Regression)

Figure 10. The Influential Observations



There are no influential points identified by Cook's distance though a lot of leverage points.

Table 5: VIF of Logit Model

	GVIF	Df	GVIF^(1/(2*Df))
Inc	1.835	1	1.355
Edu	1.400	1	1.183
Age_grp	1.612	1	1.270
state_map	1.528	50	1.004
Health_Work	1.114	1	1.055
Insure	1.161	1	1.078
Sex	1.091	1	1.044
Attitude_vacc_h	1.187	1	1.090
Attitude_h	1.169	1	1.081
Knowledge_h	1.156	1	1.075
Marriage	1.339	1	1.157
Child	1.436	1	1.198
Ill	1.093	1	1.045
Rec_h	1.018	1	1.009
employ_grp	1.473	2	1.102
MSA_grp	1.386	2	1.085

As no predictors have a VIF larger than 5, multicollinearity does not exist in our model.

Reference

1.10. decision trees. scikit. (n.d.). Retrieved April 6, 2022, from https://scikit-learn.org/stable/modules/tree.html

Baekkeskov, E. (2016). *Same Threat, Different Responses: Experts sSeering Politicians and Stakeholders in 2009 H1N1 Vaccination Policy-making.* Public Administration, 94(2), 299-315. https://doi.org/10.1111/padm.12244

Ball P. (2020). *The lightning-fast quest for COVID vaccines — and what it means for other diseases*. Nature. https://www.nature.com/articles/d41586-020-03626-1

Carolyn A. Lin & Carolyn Lagoe (2013). *Effects of News Media and Interpersonal Interactions on H1N1 Risk Perception and Vaccination Intent, Communication Research Reports*, 30:2, 127-136, DOI: 10.1080/08824096.2012.762907

Coronavirus in the U.S.: Latest Map and Case Count. The New York Times. https://www.nytimes.com/interactive/2021/us/covid-cases.html

Covid-19: First vaccine given in US as roll-out begins. BBC News. https://www.bbc.com/news/world-us-canada-55305720

E. Harrell, Frank. "*Regression Modeling Strategies*." SpringerLink, Springer, Cham, https://link.springer.com/book/10.1007/978-3-319-19425-7. 2015.

H1N1 Influenza vs. COVID-19: Pandemic Comparison. Healthline. https://www.healthline.com/health/h1n1-vs-covid-19

Henrich N, Holmes B (2011). *What the Public Was Saying about the H1N1 Vaccine: Perceptions and Issues Discussed in On-Line Comments during the 2009 H1N1 Pandemic.* PLOS ONE 6(4): e18479. https://doi.org/10.1371/journal.pone.0018479

Ledford H., Cyranoski D., Noorden V. R. (2020). *The UK has approved a COVID vaccine — here's what scientists now want to know.* Nature. https://www.nature.com/articles/d41586-020-03441-8

Mutikani L. (2021). *U.S. economy contracted 19.2% during COVID-19 pandemic recession*. Reuters. https://www.reuters.com/business/us-economy-contracted-192-during-covid-19-pandemic-recession-2021-07-29/

News Release. Bureau of Labor Statistics. U.S. Department of Labor. https://www.bls.gov/news.release/pdf/empsit.pdf

PONNAMBALAM, L., SAMAVEDHAM, L., LEE, H., & HO, C. (2012). *Understanding the socioeconomic heterogeneity in healthcare in US counties: The effect of population density, education and poverty on H1N1 pandemic mortality.* Epidemiology and Infection, 140(5), 803-813. doi:10.1017/S0950268811001464

Setbon, M., & Raude, J. (2010). Fa*ctors in vaccination intention against the pandemic influenza A/H1N1*. The European Journal of Public Health, 20(5), 490–494. https://doi.org/10.1093/eurpub/ckq054

Smith, G., Vijaykrishna, D., Bahl, J. et al. (2009). *Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic.* Nature 459, 1122–1125. https://doi.org/10.1038/nature08182

Un W., Jann W., Shan C., Yu C., Wei L., Min L., Po L., Fu H., Jen C., Yee C. (2014). *Impacts of a mass vaccination campaign against pandemic H1N1 2009 influenza in Taiwan: a time-series regression analysis*. International Journal of Infectious Diseases, Volume 23, Pages 82-89, ISSN 1201-9712. https://doi.org/10.1016/j.ijid.2014.02.016.

Vrieze, S. I. (2012), *Model selection and psychological theory: a discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)*, *Psychological Methods*. 17 (2): 228–243, doi:10.1037/a0027127, PMC 3366160, PMID 22309957