Robertson and Cao Project

Part 1:Teenage HPV Vaccination Coverage and Socioeconomic Factors

Kelly Cao and Rachel Robertson

2024-05-03

# 1. Abstract

Human Papillomavirus (HPV) is thought to be the most common sexually transmitted disease in the United States [2]. Though HPV consists of a group of over 200 related viruses, only 12 types of HPV are considered “high-risk” and associated with the incidence of six types of cancer, including oropharyngeal, cervical, penile, vulvar, vaginal, and anal cancers [4,5]. The U.S. currently recommends HPV vaccination with Gardasil 9 to immunize against six common types of high-risk HPV infections and three HPV types associated with genital warts. It is recommended that a primary dose of the HPV vaccine be given between the ages of 11 to 12 and a secondary dose administered after 6 months to 1 year from the primary [6]. Although Healthy People 2023 has set the goal of 80% HPV vaccination coverage in the U.S., only 62.6% of U.S. teenagers are up-to-date on their HPV vaccination [7]. We are interested in examining which socioeconomic and geographic factors impact HPV vaccination completion among teenagers in the U.S. To address this question, we decided to examine 2022 survey data from the National Immunization Survey for U.S. Teenagers. This survey is distributed to both providers and households of teenagers and contains a questionnaire regarding vaccination completion for several common vaccines in the U.S. immunization schedule, including HPV vaccination status. Additionally, the survey gathers information on demographics, socioeconomic status, and geographic location of the teenage patient. Using this data, we examined how several socioeconomic and demographic factors impact the status of up-to-date HPV vaccination and determined which predictors were most important in determining the up-to-date HPV vaccination status. R version 4.2.3 with R Studio was used to clean, process, explore, and analyze the survey data. Multi-step logistic regression, principal component analysis (PCA), LASSO regression, a random forest model, and an elastic net model were employed to analyze which factors are most important for determining HPV vaccination completion. The best-performing model, the random forest model, was chosen to determine the most important factors for this prediction, while the LASSO and elastic net models were used to provide a direction of association between each predictor and the outcome of HPV vaccination completion status. It was found that state is the most important factor in determining up-to-date HPV vaccination status, followed by family income, facility, age, and race or ethnicity. A high family income bracket and non-Hispanic white ethnicity were negatively associated with HPV vaccination completion while older age groups, minority ethnic groups, and lower family income were positively associated with this outcome. The relationship between state and HPV vaccination status depends on the state, which differs by legislation, cultural norms, and funding for HPV vaccination. Our results highlight the need for consistent HPV vaccination state funding and school mandates in combination with provider education to teenage patients and their parents about the efficacy and safety of the HPV vaccination.

# 2. Introduction

## 2.1 Background

Human Papillomavirus (HPV) is a double-stranded DNA virus that infects the cells of the skin and mucosal linings [1]. HPV is the most common sexually transmitted infection (STI) in the United States, where over 90% of sexually active males and 80% of sexually active females will likely become infected within their lifetime [2, 3]. Most cases of HPV are asymptomatic and cleared over time, but certain infections tend to become persistent and can cause genital warts. Specific types of HPV are known to be “high-risk” or increase the likelihood of developing cancer in persistently infected cells [4]. The HPV group contains over 200 related viruses, but only 12 types are considered “high-risk” due to the presence of oncogenes. These include HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59, though HPV-16 and HPV-18 are associated with the highest number of HPV-associated cancer cases. In the U.S., there are 37,300 cases of HPV-associated cancer cases every year, including cases of cervical, anal, oropharyngeal (throat and mouth), vulvar, vaginal, and penile cancers [5,6]. In an effort to prevent HPV-associated cancers worldwide, two HPV vaccines were developed and made available in 2006. These included a bivalent and quadrivalent vaccine that contained viral-like particles (VLPs) of the viral capsid protein (L1) from “high-risk” HPV types [7]. In recent years, the vaccine used in the U.S., Gardasil 9, has been expanded to contain HPV types 6, 11, 16, 18, 31, 33, 45, 52, and 58 [4]. Currently, the United States Centers for Disease Control and Prevention (CDC) recommends a two-dose Gardasil 9 regime to all teenagers, with a primary dose at age 11 to 12 and a second dose within 6 to 12 months of the first [8]. In 2022, the CDC estimated that 76.0% of teenagers aged 13 to 17 had received at least 1 dose of the HPV vaccine while only 62.6% of teenagers were up-to-date with their vaccination regime [9].

## 2.2 Data

The data we chose to use is the National Immunization Survey (NIS) of 2022. It consists survey data that was collected to monitor vaccination coverage for teenagers age 13-17 in the United States of America. The surveys were conducted by the National Center for Immunization and Respiratory Diseases of the Center (NCIRD) for Disease Control and Prevention (CDC). The survey itself consists of two parts: (1) the household telephone (random digit dialing) survey answered by a parent or guardian and (2) a mailed survey for the vaccination provider, called “Immunization History Questionnaire.” The original data set is a fixed width file (FWF). In the data set, there are 43,738 rows and 672 columns, with each row representing an individual and each column representing the answer to a question. The questions pertain to the teenager’s immunization history, demographics, and additional household-reported health information. We have truncated the data set to include 26-selected columns for the ease of observation. We have also filtered the data set to include one year, 2022, which is the most recent year of data that is published. The provided survey data will require cleaning, as there are signs of human errors present.

## 2.3 Research Question

Data from the 2022 National Immunization Survey for teens will be used to identify which socioeconomic and geographic factors are associated with HPV vaccination completion for teenagers aged 13 to 17 in the U.S. Specifically, we want to find the most important socioeconomic and geographic factors for a teenager being up-to-date for their HPV vaccinations in the U.S. The outcome of this study will be measured using the overall up-to-date HPV vaccination completion indicator, which includes all age groups in any step of the vaccination process. The variable includes teenagers aged 13 to 17 who are up-to-date and have received at least one HPV shot.

## 2.4 Purpose

Measuring HPV vaccine completion with “up-to-date” variables may be more robust than vaccination rate based on those who have completed the regime (2 or more shots), because not all teens included in the survey are at the age in which they would’ve completed the full regime. This means that younger teens who are up to date with 1+ shot will not be excluded from the analysis. We aim to measure determinants of socioeconomic status that are engrained into the survey questions. These factors include: family income, poverty status, income-to-poverty ratio, insurance status, insurance breaks, maternal education, living arrangement, and geographic mobility status. Additional demographic factors that are associated with healthcare access and may be examined include: race, ethnicity, language, facility in which the vaccine was administered, and whether the teen had completed a wellness exam between the ages of 11-12. Geographic distribution of the teens will be assessed by true state of residence. We would like to examine the data for correlations between HPV vaccination completion status and the aforementioned socioeconomic determinants, demographics, and geographic location. This serves as valuable information to determine differences in healthcare access and vaccination coverage for teenagers living in the U.S.

It is generally known that those of lower income and without insurance have limited healthcare access in the U.S. [10], but it will be interesting to see if this pattern is reflected at the state level. If so, policy should be examined as it may be related to healthcare access disparities. Recent studies have examined HPV vaccination completion according to social determinants, but they have not examined geographic distribution and include data prior to 2019 [11]. More recent years display differences in global HPV vaccination trends due to the COVID19 pandemic [12]. Our study will fill in geographical gaps and provide updated vaccination trends based on socioeconomic factors.

# 3. Methods

## 3.1 Loading Data

The data was loaded from the original DAT file, entitled ‘NISTEENPUF22.DAT’ by specifying the column positions within the data file. This is performed with the readr library and its fwf\_position function. The columns were then read and made into a data frame. Each of the factor variables was defined and the corresponding levels were assigned labels, based on the codebook. This is performed with the dplyr library and its mutate function. The structure and summary statistics of all of the variables were examined. The loaded data is saved as a rds and a csv in the ‘processed-data’ folder, as to not alter the original raw data file. The definition of each factor variable in included in the README file within the ‘raw-data’ folder.

## 3.2 Cleaning Data

The structure of the processed data was examined. The data was examined for missing values. No NAs were present, however missing values were present as labeled factor levels. The missing factor levels were printed, identified, and replaced with NA values. A string of 27,000 NAs were present in the outcome variable, indicating vaccination completion. All rows corresponding to the NAs for the missing outcome value were removed. Approximately 160 additional rows with NA values for facility or insurance status variables were removed. Over 16,000 rows of data remained. The year column was eliminated, as all of the data is from 2022. We saved the cleaned data as an rds and csv in the ‘processed-data’ folder.

## 3.3 Exploratory Analysis

Exploratory analysis was used to determine the distribution of the predictors, including the socioeconomic variables that were chosen for analysis. The distribution of single predictors including income-to-poverty ratio, income, state, race, ethnicity, vaccination status, and facility were examined using histograms and scatter plots. The following predictors were compared visually and fitted to a linear regression to examine preexisting associations: vaccination status and ethnicity, vaccination status and poverty status, income poverty ratio and insurance status, income poverty ratio and ethnicity.

## 3.4 Analysis

### 3.4.1 Generalized Logistic Regression

A generalized logistic regression was performed with bi-directional step-wise selection for predictor reduction. Variables that were redundant of a major variable were excluded from analysis. The variables initially included were age, sex, state, race/ethnicity, insurance status, income, insurance breaks, poverty status, education, language, geographic mobility, facility, and housing status. The model was evaluated with a residual plot and correlation matrix to reveal strong multicollinearity. This is the basis for the subsequent analysis methods, as they are more resistant to collinearity.

### 3.4.2 Principal Component Analysis

Principal component analysis (PCA) was performed to do dimensionality reduction. PCA would ideally transform the highly correlated predictors into uncorrelated components. The h2o package was used to provide a machine-learning platform in RStudio, to allow machine-learning methods ot be implemented. Due to the data being mainly categorical variables, the PCA proved to be unsuccessful. A LASSO Ridge Regression was performed with the h2o package, to mitigate the multicollinearity issue.

### 3.4.3 Random Forest Model

A random forest model was used, as it is resistant to multicollinearity. The tidymodels package was used to produce a random forest model with the classification method. The model was trained and fit from training data, which was an 80% split of the original data. After evaluation, the model parameters for minimal node size and number of selected predictors were tuned for using a grid search with cross-validation resamples. After tuning, the best model was extracted based on the ROC AUC metric and was fit to the training data. The tuned model was used to make predictions and evaluated with performance metrics, a confusion matrix, and a comparison to a single-predictor random forest model. Finally, the tuned random forest model was re-evaluated by being fit to the testing data.

### 3.4.4 LASSO Regression

A LASSO model was produced as it is resistant to multicollinearity because it implements regularization and variable selection, by reducing unimportant coefficients to zero. Tidymodels in R was used to produce a logistic regression LASSO model using the glmnet engine. The model recipe included all of the selected predictors from the stepwise selection and was fit to the training data split from 80% of the original data. The penalty parameter was tuned from a grid using cross-validation and the ideal penalty was selected using the model that maximized ROC AUC. The tuned model was extracted and re-fit to the training data. The metrics accuracy, F1, and confusion matrix were compared between tuned and original fit models. The tuned model was fit to the testing data and the most important variables were extracted using the vip package.

### 3.4.5 Elastic Net Regression

An elastic net regression was produced using the tidymodels framework in R using the glmnet engine set to classification mode. The initial elastic net model included all of the variables selected by the step-wise selection above. The model was then tuned for the parameters penalty and mixture using a cross-validation grid search. The ideal model was selected by maximizing ROC AUC for both of the tuning parameters. The final model was fit to the test data and re-evaluated with the metrics: accuracy and kappa. The most important variables were extracted using the vip package.

# 4. Results

## 4.1 Exploratory Data Analysis

Exploratory analysis is used to determine the distribution of our predictors, including the socioeconomic variables that we chose for analysis. Nearly all of the demographic factors have a similar distribution between those who have completed their HPV vaccination regime versus those who have not completed it.

#| label: Table1   
  
#| tbl-cap: "Table 1.Socioeconomic characteristics of U.S. teenagers who have completed or are up-to-date with HPV vaccination series."   
  
#| echo: FALSE   
  
Table1 <- readRDS(here("results/tables/table1.rds"))   
  
kable(Table1)

| **Characteristic** | **NOT UTD**, N = 3,465 | **UTD**, N = 13,099 | **Missing Data**, N = 0 |
| --- | --- | --- | --- |
| Age (years) | NA | NA | NA |
| 13 | 845 (24%) | 2,448 (19%) | 0 (NA%) |
| 14 | 753 (22%) | 2,757 (21%) | 0 (NA%) |
| 15 | 657 (19%) | 2,667 (20%) | 0 (NA%) |
| 16 | 603 (17%) | 2,707 (21%) | 0 (NA%) |
| 17 | 607 (18%) | 2,520 (19%) | 0 (NA%) |
| Sex | NA | NA | NA |
| MALE | 1,945 (56%) | 6,741 (51%) | 0 (NA%) |
| FEMALE | 1,520 (44%) | 6,358 (49%) | 0 (NA%) |
| Insurance Status | NA | NA | NA |
| PRIVATE INSURANCE ONLY | 2,020 (58%) | 7,682 (59%) | 0 (NA%) |
| ANY MEDICAID | 1,025 (30%) | 4,205 (32%) | 0 (NA%) |
| OTHER INSURANCE (CHIP, IHS, MILITARY, OR OTHER, ALONE OR IN COMB. WITH PRIVATE INSURANCE) | 305 (8.8%) | 968 (7.4%) | 0 (NA%) |
| UNINSURED | 115 (3.3%) | 244 (1.9%) | 0 (NA%) |
| Poverty Status | NA | NA | NA |
| ABOVE POVERTY > $75K | 1,930 (56%) | 7,415 (57%) | 0 (NA%) |
| ABOVE POVERTY <= $75K | 1,063 (31%) | 3,523 (27%) | 0 (NA%) |
| BELOW POVERTY | 370 (11%) | 1,814 (14%) | 0 (NA%) |
| UNKNOWN | 102 (2.9%) | 347 (2.6%) | 0 (NA%) |
| Race and Ethnicity | NA | NA | NA |
| HISPANIC | 575 (17%) | 2,728 (21%) | 0 (NA%) |
| NON-HISPANIC WHITE ONLY | 2,224 (64%) | 7,514 (57%) | 0 (NA%) |
| NON-HISPANIC BLACK ONLY | 298 (8.6%) | 1,218 (9.3%) | 0 (NA%) |
| NON-HISPANIC OTHER + MULTIPLE RACE | 368 (11%) | 1,639 (13%) | 0 (NA%) |
| Maternal Education-level | NA | NA | NA |
| LESS THAN 12 YEARS | 165 (4.8%) | 782 (6.0%) | 0 (NA%) |
| 12 YEARS | 508 (15%) | 1,943 (15%) | 0 (NA%) |
| MORE THAN 12 YEARS, NON-COLLEGE GRAD | 1,047 (30%) | 3,095 (24%) | 0 (NA%) |
| COLLEGE GRADUATE | 1,745 (50%) | 7,279 (56%) | 0 (NA%) |
| Survey Language | NA | NA | NA |
| ENGLISH | 3,297 (95%) | 11,993 (92%) | 0 (NA%) |
| SPANISH | 148 (4.3%) | 1,019 (7.8%) | 0 (NA%) |
| OTHER | 20 (0.6%) | 87 (0.7%) | 0 (NA%) |
| Housing Status | NA | NA | NA |
| OWNED OR BEING BOUGHT | 2,622 (76%) | 9,810 (75%) | 0 (NA%) |
| RENTED | 737 (21%) | 2,951 (23%) | 0 (NA%) |
| OTHER ARRANGMENT | 98 (2.8%) | 312 (2.4%) | 0 (NA%) |
| DON’T KNOW | 4 (0.1%) | 8 (<0.1%) | 0 (NA%) |
| REFUSED | 4 (0.1%) | 18 (0.1%) | 0 (NA%) |

Income-poverty-ratio is a numeric value that indicates how close to poverty a family is living. It is calculated by income over the poverty income threshold and ranges from values 1 (closest to poverty) and 3 (furthest from poverty). The distribution of the income poverty ratio indicates a strong left skew. The highest frequency of observations was gathered from households with an income poverty ratio, meaning most survey participants live well above the poverty threshold. The family income is right skewed with the majority of the observations coming from $75000+ income level. This indicates that a high frequency of participants are above the poverty income threshold, but also in a modest income class.

|  |
| --- |
| Figure 1. Income-poverty ratio distribution |

|  |
| --- |
| Figure 2. The family income distribution |

The state variable shows a more uniform distribution, suggesting that the observations were gathered fairly across the US states and territories. States and territories with a significantly larger number of observations include Texas, West Virginia, Pennsyvania, Puerto Rico, New York, and Illinois. Although it is noted a few territories provide 0 observations. The race and ethnicity variables show a distribution that is similar to the U.S. census data. The census data states that the racial composition of the U.S. is 58.9% white, 13.6% African American, and 19.1% Hispanic or Latino, while our survey population is composed of 58.% white, 9.15% African American, and 19.9% Hispanic or Latino participants.

The majority of observations came from households with private insurance only The Insurance status variable shows that the majority of the observations came from households with private insurance only (58.6%), followed by Medicaid (31.6%), other insurance (7.7%), and uninsured (2.2%). This is mirrored in the facility variable, as survey location, where a majority of its observations show private facilities (34.4%), followed by mixed facilities (25.8%), hospitals (11.3%), and all public facilities (9.9%). 15.2% of the facility types were unknown. Upon stratification of facility type by HPV vaccination completion status, it was found that the majority of participants who were up-to-date were surveyed at private or mixed facilities. They likely received their HPV vaccination there as well.

|  |
| --- |
| Figure 3.Facility and vaccine completion distribution |

A total of 79% of participants had up-to-date status for receiving 1 or more shots of HPV vaccination. U.S. states and territories had an HPV vaccination completion rate ranging from 69% to 95%, fluctuating around ~80%. There are different counts of each race/ethnicity for each factor level, so the percentage of those with up-to-date vaccine status out of each race/ethnicity level was adjusted. The Hispanic ethnicity was found to have the highest vaccine completion while the non-Hispanic white ethnicity had the lowest. It was also found that for Hispanic and Non-Hispanic black ethnicities, the income poverty ratio was higher for those who were not up-to-date in HPV vaccination status, compared to those who were. The opposite was true for mixed ethnic groups, where vaccine completion corresponded with a higher mean income-poverty ratio. When examining the overall HPV vaccination status stratified by income poverty ratio, the income poverty ratio did not differ between those who are up to date and those who are not. This may be accounted for by other trends such as that shown with ethnicity, above.

|  |
| --- |
| Figure 4. Distribution of income poverty ratio stratified by race and HPV vaccination completion status |

Subsequently, income poverty ratio and insurance status were examined, and stratified by HPV vaccination status. It is apparent that those with private and other insurance status had the highest income-poverty ratio, while those with Medicaid or uninsured were closer to the poverty income threshold. For those with Medicaid, those who were up-to-date in HPV vaccinations had a slightly lower income-poverty ratio than those who were not up to date.

|  |
| --- |
| Figure 5. Income poverty ratio and insurance status |

## 4.2 Analysis

The variables selected for analysis were chosen based on their interpretability and applicability to our research question “Which socioeconomic and geographic factors impact HPV vaccine completion rates among teenagers?” Any variables that capture repetitive data were removed and the most important or straightforward version of the variable was kept in the model. Seven variables represented data for the outcome of HPV vaccination completion:

|  |
| --- |
| Table 2. Human Papillomavirus vaccine status variables and meanings. |

The first option, P\_UTDHPV was the chosen outcome as it is the most general. It has the benefit of being able to use Age as a predictor variable and also not biasing the model towards older ages who are required to receive more than one shot to be considered up-to-date. Additionally, in April of 2022, the SAGE work group of the WHO officially recommended a single-dose HPV vaccine regime, because vaccine efficacy was found to be nearly as high with a single dose as it is with the multi-dose regime. This recommendation was given with the goal of increasing vaccine coverage in areas with lower vaccine accessibility and availability (13). In September of 2023, PAHO followed suit in recommending a single-dose HPV vaccine regime for the Americas (14). This supports our rationale for choosing the single or more dose, up-to-date, outcome variable. The variables that remained in the initial logistic regression are listed below.

|  |
| --- |
| Table 3. Potential predictor variable names and meanings. |

INCOPAR or income-poverty ratio was excluded because this is reflected by the income and poverty status variables, separately. Additionally, this is a numeric variable while the others are categorical. RACE\_K or race was excluded because race is included within the RACEETHK variable. RACEETHK is ideal because it includes both race and ethnicity.

### 4.2.1 Stepwise Selection with Logistic Regression

To determine which predictor variables will be best suited for a general logistic model, we employed a bi-directional stepwise selection method. The results were assessed with AIC and WAIC. The stepwise model is set up with P\_UTDPHPV as the outcome and the remaining chosen variables as the predictors. Each predictors were added and dropped in different variations, and the resulting AIC were recorded for each of the different models. From this, we were able to determine that the model with the lowest AIC (and thus the greatest fit) was the model that included all of the original factors.

The logistic regression with the selected variables was assessed by plotting the residuals. This revealed that the predictors had a high indication of collinearity. To confirm this, a VIF analysis was attempted but was unable to proceed as the multicollinearity was too great for too many variables. As a result, we decided to perform a correlation matrix, which performed as expected and verified our concerns. Multi-collinearity was present between the majority of the predictor variables, thus another model type that could handle collinearity would be ideal.

|  |
| --- |
| Figure 6. Residual Plot Results |

### 4.2.2 Principal Component Analysis and LASSO Regression

In order to reduce the multicollinearity issue, we initially attempted to use PCA. To employ the unsupervised learning model, we use the h2o package. The PCA resulted in one principle component with a high standard deviation and high proportion of variance, showing that the Pc1 captures a large portion (97.9%) of variability in the data.

|  |
| --- |
| Table 4. Summary of PCA Attempt |

The result of having only one principle component further supports the fact that many of the predictor variables are highly collinear. One PC shows that a PCA must reduce the data object to a single dimension. Having a single principal component could be problematic for my analysis, particularly if it leads to the loss of crucial information or structure in my data due to collinearity among the original features.

Using the same package, I used the LASSO logistical regression model and performed a co-validation test on it.

|  |
| --- |
| Table 5. Summary of Initial LASSO Regression |

### 4.2.3 Random Forest Model

We used a Random Forest Model to assess which predictors are associated with HPV vaccination completion rates in U.S. teens. Decision tree models are known to be able to handle collinearity and predictors with different trends. A random forest model is ideal because it employs bootstrapping to produce randomly chosen iterations of decision trees, which introduces variability into the models. The best random forest model can then be tuned for and selected based on which tree performs best in a certain metric. The random forest model was produced using the tidy model framework and tuned using a cross-validated grid search for the minimum node size (min\_n) and the number of variables to randomly sample at each split (mtry). The model with the best ROC AUC metric was selected, and predictions were produced. This model initially indicated good performance with accuracy and F1 measure of 0.95 and 0.86 respectively. A confusion matrix estimated an out-of-bag (OOB) error rate of 23.0%, a specificity of 99.4%, and a sensitivity of 76.6%. The most important predictors for HPV vaccination completion from this model, in order of most to least important, were state, income level, facility, age, race/ ethnicity, insurance status, education, housing status, sex, and geographic mobility.

|  |
| --- |
| Figure 7. Top Predictors Chosen by Random Forest Model |

However, after testing the same metrics with new data (testing data), the F1 metric was reduced from 0.86 to 0.16 and the accuracy decreased from 0.95 to 0.78, which could indicate overfitting or that the model is “too big” for the data. To address the concern of overfitting, a LASSO model was utilized.

#### 4.2.3.1 LASSO Regression with Tidymodels

The LASSO regression was performed using the Tidymodels framework. A LASSO regression is ideal for multicollinearity as it not only regularizes the variables but also performs variable selection by reducing the coefficients of unimportant variables to zero. The LASSO model was tuned for the penalty parameter and selected based on the best ROC AUC performance. The tuned model improved in performance from the initial model, with an initial ROC AUC of 0.37 and a tuned ROC AUC of 0.34. The accuracy of the initial and tuned model remained the same, at 0.79. The F1 measure increased from 0.03 to 0.07 from the initial to the tuned model. The specificity and sensitivity of the tuned model were calculated from the confusion matrix, to be 98.7% and 3.4%, respectively. This means the model is highly likely to produce false negatives or predict that HPV vaccination was not completed when in reality it was. The most important predictors, extracted from the tuned model from most to least important, were state (Rhode Island), facility (STD, school, or teen clinics), state (Vermont), state (Mississippi), state (North Dakota), state (Iowa), state (Hawaii), housing status (unknown), state (Nebraska), and insurance breaks (currently uninsured or never insured since 11 years of age).

|  |
| --- |
| Figure 8. Top Predictors chosen by LASSO Regression Model |

#### 4.2.3.2 Elastic Net Regression

In an attempt to produce a better-performing model compared to the LASSO regression, an elastic net regression was produced using the Tidymodels framework. The elastic model employs methods from both the LASSO regression and Ridge regression by implementing a penalty parameter and mixture parameter. The model was fit initially and then tuned for both the penalty and mixture parameters using cross-validation to maximize ROC AUC. The initial fit elastic model and tuned model both performed with an ROC AUC of 0.34. The tuned model had an accuracy and kappa measure of 0.79 and 0.06, respectively. After fitting the tuned model to testing data, the accuracy and kappa were reduced to 0.79 and 0.4 respectively. The ROC AUC for this fit increased to 0.36 and the F1 measure was 0.07. The performance of the elastic net model was very similar to the LASSO model. Neither the LASSO nor elastic net model performed as well as the random forest tree model, despite the overfitting of the random forest model. The most important predictors chosen by the elastic net model were the same as the predictors chosen by the LASSO regression. The most important predictors in determining HPV vaccination completion, from most to least important, were state (Rhode Island), facility (std, school, and teen clinics), state (Vermont), state (Mississippi), state (North Dakota), state (Iowa), state (Hawaii), housing status (unknown), insurance breaks (currently uninsured or never insured since 11 years of age), and state (Nebraska).

# 5. Discussion

The random forest model was chosen to predict HPV vaccination completion, as it was the best-performing model for all of the performance metrics, despite potential overfitting to the data. The other two machine learning models, the LASSO and elastic net models, may also provide insight into which predictors are most important for determining HPV vaccination completion. These models are mainly in agreement with each other and display similar variable importance rankings as the random forest model. The findings will be interpreted by examining which variables were chosen by all three models, as they may hold the most importance, followed by the predictors chosen by the random forest model.

For all three model types that were tested, state was chosen to be the most important factor in determining HPV vaccine completion status. Both the LASSO regression and elastic net models determined that Rhode Island was the state most closely associated with HPV vaccine completion, followed by Vermont, Mississippi, North Dakota, Iowa, Hawaii, and lastly, Nebraska. For both of these models, all of these states have positive coefficients associated with the outcome of up-to-date HPV vaccination, except for Mississippi, which has a negative coefficient. This may be explained by the fact that each state determines HPV vaccination legislation, policy, and funding, for teenagers aged 11 to 13. As of 2021, four states have a law requiring vaccination for school entry, including,  Hawaii, Rhode Island, Virginia, and Washington D.C (15). As of 2019, four states recommend HPV vaccination for school attendance, by law, including, Arkansas, Delaware, West Virginia, and Wyoming (16). As of 2019, 42 U.S. states and jurisdictions provide legislative funding to promote HPV vaccination, however, Rhode Island is the only state to implement both a school entry requirement and funding for HPV vaccination (17). Additionally, Rhode Island has the highest HPV vaccination coverage out of all U.S. states and jurisdictions, exceeding 80% coverage as of 2022 (18). The states selected as most important by the models that have legislation to fund HPV vaccination campaigns include Rhode Island, North Dakota, Mississippi, and Iowa (19). States chosen as important by the model with legislation mandating HPV vaccination are Rhode Island and Hawaii. Neither Vermont nor Nebraska have legislature funding and mandating HPV vaccination, although, both are in the top 15 states for HPV vaccination coverage in the U.S. (18,19). Conversely, Mississippi has a legislature to reimburse the cost of HPV vaccination to insurance providers but has the lowest HPV vaccination coverage in the country. This is thought to be due to misconceptions and misinformation about the HPV vaccine’s purpose and the way that HPV is transmitted (18,20). The mixture of policies and vaccination coverage rate, which is not always policy dependent, demonstrates the importance of a combination of legislation to mandate and fund HPV vaccination as well as provider and patient education to support HPV vaccination.

All three models also indicated that facility and housing status are important in determining up-to-date HPV vaccination status. According to the LASSO and elastic net models, the facility being an STD clinic, school clinic, or teen clinic, is negatively associated with HPV vaccination completion, while all other facility categories (hospitals, private facilities, mixed facilities) are positively associated with the up-to-date HPV vaccination status. This may indicate that STD clinics, school clinics, and/or teen clinics neglect to recommend or educate their teenage patients on HPV vaccination. Although school-based HPV vaccination programs have shown promise in increasing HPV vaccine uptakes internationally, these programs in the U.S. face barriers such as politicization of vaccines, patient consent processes for teens, and funding for these programs (21,22). An unknown housing status was also negatively associated with the outcome of up-to-date HPV vaccination in the LASSO and elastic net models. This may reflect that housing security indirectly impacts HPV vaccine uptake in teenagers. A cross-sectional study and analysis of house-based socioeconomic status index in the U.S. in 2020, indicated that HPV vaccination initiation was related to housing and socioeconomic status, where lower quartiles corresponded to lower rates of HPV vaccine initiation (23).

Along with the variables state and facility, the random forest model indicated family income, age, and race/ ethnicity to be in the top five most important predictors for HPV vaccination completion. The LASSO and elastic net models both indicate that family income at the highest three income brackets is negatively associated with HPV vaccination completion, while the lowest three income brackets are positively associated with HPV vaccination completion. In contrast to our findings, a CDC data brief found that the percentage of teenagers who received one or more HPV vaccine doses increased as family income moved to be greater than the federal poverty level. However, this study does not differentiate vaccination levels at the uppermost income brackets (24). A recent study analyzing National Immunization Teen Surveys from 2017 to 2021 found that 64.7% of parents in advantaged income brackets had no intention to begin HPV vaccination while 40.9% of parents in disadvantaged income brackets had no intention to begin the vaccine regime. Additionally, the advantaged group was less likely to complete HPV vaccination than the disadvantaged group. For the advantaged group, the main reason for not starting the regime was safety concerns and the main reason for not completing the regime was the belief that their child was up-to-date on HPV vaccinations (25). This highlights the importance of addressing vaccination misinformation and provider education to parents on the recommended HPV vaccination schedule.

Age was selected as the fourth most important predictor of HPV vaccination completion by the random forest model, but not selected by the LASSO or elastic net models, so the direction of association for this study is unknown. A 2024 data brief by the CDC found that the percentage of teenagers who have received at least one dose of the HPV vaccination increases with age, despite the recommended age for beginning the HPV vaccination being 11 to 12 years old in the U.S. (24). Next, Race and ethnicity were selected as the fifth most important predictor of HPV vaccination completion by the random forest model, and selected by the LASSO and elastic net models, though it was not found to be the top most important predictor by this model. The LASSO and elastic net models found that being part of the non-Hispanic ethnic group is negatively associated with the up-to-date HPV vaccination status, while all other racial and ethnic groups are positively associated with the up-to-date HPV vaccination status. Current literature provides mixed results on the association between race and ethnicity and HPV vaccination status. Several studies found that Hispanic children are less likely to have received one or more HPV vaccination doses than non-Hispanic white children, while other studies indicate that vaccine initiation is lower is non-Hispanic whites than in racial and ethnic minority groups. However, these resources support that HPV vaccine follow-through may be lower in racial and ethnic minority groups, compared to lower HPV vaccine initiation rates in non-Hispanic white teenagers (24,26,27,28). Overall, larger-scale, in-depth studies are necessary to determine the association between racial and ethnic group differences in HPV vaccination initiation and completion rates in the U.S.

# 6. Conclusion

The analysis studying the socioeconomic factors impacting HPV vaccination completion rates among teenagers yielded many valuable insights despite some challenges that were involved.

The exploratory data analysis (EDA) aimed to understand the distribution of potential predictors and their impact on vaccination status. The combination of descriptive statistics, visualizations, and statistical modeling provided a comprehensive understanding of the data set.

Subsequent attempts to fit the data to models met with some challenges and limitation. The Logistic Regression models encountered multicollinearity issues, which prompted the use of subsequent techniques/tools like PCA, LASSO regression, or even elastic net regression. The PCA led to the reduction of only one principle component, which made it impossible for subsequent analysis.

While the LASSO regression model and the elastic net regression model attempt at solving the multicollinearity issue was satisfactory, both proved itself to be a poor model with low predictive power. The Random Forest model was initially promising, but ultimately showed signs of overfitting, thus limiting its reliability.

There are likely other factors beyond those captured in the dataset, such as cultural norms, influencing HPV vaccination completion rates among teenagers in the US. These factors could not be adequately captured by the trends identified in the analysis.

In addressing collinearity concerns, feature selection played a pivotal role during the variable selection process, ensuring that only non-repetitive variables were chosen. This deliberate strategy aimed to mitigate collinearity effects and enhance the robustness of our analysis. Despite these efforts, however, several barriers, including model limitations and logistical challenges, persisted. Moving forward, refining modeling techniques and embracing collaborative approaches remain essential for advancing our understanding and overcoming these challenges.

# 7. References

1. https://www.who.int/teams/health-product-policy-and-standards/standards-and-specifications/vaccine-standardization/human-papillomavirus#:~:text=Human%20papillomavirus%20(HPV)%20is%20a,approximately%208%2Dkb%20in%20length.
2. https://www.womenshealth.gov/a-z-topics/human-papillomavirus#:~:text=Genital%20HPV%20is%20the%20most,79%20million%20Americans%20have%20HPV.&text=It%20is%20so%20common%20that,some%20point%20in%20their%20lifetime.
3. https://www.health.state.mn.us/data/mcrs/data/qfhpv.html#:~:text=Human%20Papillomavirus%20(HPV)%20is%20a,with%20HPV%20in%20their%20lifetime.&text=Around%2050%20percent%20of%20HPV,HPV%2C%20which%20can%20cause%20cancer.
4. HPV and Cancer. In: National Cancer Institute [Internet]. 1 Mar 2019 [cited 12 Mar 2024]. Available: https://www.cancer.gov/about-cancer/causes-prevention/risk/infectious-agents/hpv-and-cancer
5. https://www.cdc.gov/vaccines/vpd/hpv/hcp/vaccines.html#:~:text=Vaccine%20Composition,45%2C%2052%2C%20and%2058.
6. https://www.cdc.gov/vaccines/vpd/hpv/hcp/recommendations.html#:~:text=Two%20doses%20of%20HPV%20vaccine,third%20dose%20of%20HPV%20vaccine.
7. https://www.cdc.gov/mmwr/volumes/72/wr/mm7234a3.htm#:~:text=In%202022%2C%2089.9%25%20of%20adolescents,HPV%20vaccination%20(HPV%20UTD).
8. Sciences, N. A. of, Engineering, & Medicine. (2018). Factors that affect health-care utilization. Health-Care Utilization as a Proxy in Disability Determination. https://www.ncbi.nlm.nih.gov/books/NBK500097/
9. Mansfield, L. N., Chung, R. J., Silva, S. G., Merwin, E. I., & Gonzalez-Guarda, R. M. (2022). Social determinants of human papillomavirus vaccine series completion among U.S. adolescents: A mixed-methods study. SSM - population health, 18, 101082. https://doi.org/10.1016/j.ssmph.2022.101082
10. Casey, R. M., Akaba, H., Hyde, T. B, et al. (2024). Covid-19 pandemic and equity of global human papillomavirus vaccination: descriptive study of World Health Organization-Unicef vaccination coverage estimates. BMJ Medicine, 3, 000726. doi: 10.1136/bmjmed-2023-000726
11. https://www.who.int/news/item/20-12-2022-WHO-updates-recommendations-on-HPV-vaccination-schedule
12. https://www.paho.org/en/news/5-9-2023-paho-technical-advisory-group-recommends-countries-americas-use-single-dose-hpv
13. https://www.kff.org/womens-health-policy/fact-sheet/the-hpv-vaccine-access-and-use-in-the-u-s/
14. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6746494/
15. https://hpvroundtable.org/wp-content/uploads/2019/09/05\_School\_Entry\_Requirements\_for\_HPV\_Vaccination\_WEB\_updated.pdf
16. https://chlpi.org/wp-content/uploads/2022/07/CHPLI-HPV-Vax-design-v9-web.pdf
17. https://mississippitoday.org/2023/08/21/hpv-vaccine-mississippi/
18. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10478734/
19. https://pubmed.ncbi.nlm.nih.gov/30735468/
20. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7484398/
21. https://www.cdc.gov/nchs/products/databriefs/db495.htm#:~:text=degree%20or%20higher.-,The%20percentage%20of%20children%20who%20have%20received%20one%20or%20more,of%20the%20federal%20poverty%20level.
22. https://www.thelancet.com/journals/lanam/article/PIIS2667-193X(24)00021-8/fulltext
23. https://www.tandfonline.com/doi/full/10.1080/21645515.2024.2313249#:~:text=With%20regard%20to%20race%2Fethnicity,likely%20to%20have%20been%20vaccinated.
24. https://pubmed.ncbi.nlm.nih.gov/29347831/
25. https://www.sciencedirect.com/science/article/abs/pii/S0091743519301148