Robertson and Cao Project

Part 1:Teenage HPV Vaccination Coverage and Socioeconomic Factors

Kelly Cao and Rachel Robertson

2024-03-16

# 1. Introduction

## 1.1 Data Set

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.

## 1.2 Research Question

We want to use the 2022 National Immunization Survey for teens to identify which socioeconomic and geographic factors are associated with HPV vaccination completion for teenagers in the U.S. Specifically, we are asking the question, What is the likelihood of a teenager being up-to-date for their HPV vaccinations in the U.S., based on socioeconomic factors and geographic distribution. The outcome of this study will be measured using the columns that pertain to HPV vaccination completion. These variables are provider-collected, and they state whether or not the teen is (1) up-to-date with 1+ HPV shot, (2) up-to-date with 2+ HPV shots, or (3) up-to-date with 3+ HPV shots (excluding all vaccinations post-survey). 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.

## 1.3 Proposed Analysis

For the analysis portion, we have decided to do a correlation analysis between HPV vaccination completion and each of the socioeconomic factors. When performing this, we may choose to stratify by age or gender, if we find these to be confounders. The correlation analysis cannot be performed using a multivariate model because there may be collinearity present between many of the determinants for socioeconomic status. This will lead to spurious results. Instead, the correlation between vaccination status and each factor is examined separately. Next, we may stratify this by state. Comparing these correlation values between states and regions may lead us to find patterns of healthcare accessibility in the U.S. It is generally known that those of lower income and without insurance have limited healthcare access in the U.S.(1), 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 (2). More recent years display differences in global HPV vaccination trends due to the COVID19 pandemic (3). Our study will fill in geographical gaps and provide updated vaccination trends based on socioeconomic factors.

# 2. Methods

## 2.1 Data Cleaning and Processing

### 2.1.1 Loading the data

The data was loaded from the original DAT file, entitled ‘NISTEENPUF22.DAT’ by specifying the column positions within the data file. The columns were then read and made into a data frame. Next, each of the factor variables were defined and the corresponding levels were assigned labels. 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. This does not alter the original raw data file. The definition of each factor variable in included in the README file within the ‘raw-data’ folder.

Next, we explored the structure of the processed data and searched for any missing values. However, there were no NAs present. There are, however, missing values as labelled factor levels which must be accounted for. We identified the missing value labels by printing the factor levels. We then replaced the all of the values labelled missing with NA values. Next, we found a string of over 27000 NAs all of the variables for vaccination completion. These rows corresponded for all variables related to vaccine completion. Since this is the response variable of our scientific question, we eliminated these rows. We also eliminated rows with small numbers missing values, which were within the columns for facility, Wellchild exams, and insurance breaks. Lastly, we eliminatethe year column, as all of the data comes from the year 2022. We saved the cleaned data as an rds and csv in the ‘processed-data’ folder.

## 2.2 Exploratory Analysis

First we loaded the packages needed for the exploratory analysis. And then we loaded the data called ‘cleandata1’ from the ‘processed-data’ folder.

We start our exploratory analysis by displaying the distribution of our predictors, including the socioeconomic variables that we chose for analysis. The distribution of the numerical income to poverty ratio can be shown below. It displays a value from 1-3. A ratio displaying less than 1 depicts an income less than the poverty level.Any income ratio >1 indicates an income greater than the poverty level. Most of the observations were gathered from households with income ratios of 3, as shown by the skewed plot below.

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

Nest, the distribution of surveys given per state was analyzed. The state variable shows a more uniformed distribution, suggesting that the observations were gathered fairly across the US states and territory. Although it is noted a few territories provides 0 observations. Figure 2. State dsitribution

The family income variable provides a more normal distribution with the majority of the observation coming from $75000+ income level. Figure 3. The family income distribution

The race and ethnicity variables show a distribution that is similar to the U.S. census data, which concludes that the racial composition of the U.S. is 58.9% white, 13.6% African American, and 19.1% Hispanic or Latino.

Figure 4.Racial distribution Figure 5.Ethnicity distribution

The Insurance status variable shows that the majority of the observations came from households with private insurance only. This is followed by any medicaid and then other insurance. This is mirrored in the facility variable, where a majority of its observations show private facilities being the location. Figure 6. Insurance status distribution Figure 7. Distribution of the medical facility where the survey was administerred

The vaccination variable indicates nearly 80% of the observations had an Up-to-date status for the 1+ shot HPV vaccination. We will follow by examining the percentage of up-to-date vaccination status by state.

|  |
| --- |
| Figure 8. Vaccine status distribution |

We examine the facility that submitted the survey, which is stratified by the vaccine completion status of the survey subject.

|  |
| --- |
| Figure 9.Facility and vaccine completion distribution |

We examine the percentage of vaccination completion by race and ethnicity. Because there are different counts of each race/ethnicity for each factor level, the percentage of those with up-to-date vaccine status out of each race/ethnicity level was determined. This adjusts for the different counts in each racial category. The Hispanic ethnicity was found to have the highest vaccine completion while non-hispanic white has the lowest.

|  |
| --- |
| Figure 10. Distribution of vaccine completion stratified by race and ethnicity |

## 2.3 Model and Plot Fitting

A boxplot is created with income-poverty ratio to vaccination status. The box plot indicates little difference between the two status. A statistical model is fitted to the same variables. A p-value of 0.1358 suggests there is little significance between the income-poverty variable and the vaccination status.

|  |
| --- |
| Figure 11. |

A boxplot is created with income-poverty ratio to insurance status. The box plot indicates there are distinct difference between the two status. A statistical model is fitted to the same variables. The p-value of < 2.26e-16 suggests there is statistical significance between the income-poverty variable and the insurance status.

|  |
| --- |
| Figure 12. Income poverty ratio and insurance status |

Income-poverty ratio was fit to vaccine status and insurance status in a multi-linear model.

A box-plot was created for income-poverty level versus race/ethnicity and this was stratified by vaccine completion (UTD or NOT UTD). A lower income-poverty ratio means the group is closer to the poverty line. For the non-Hispanic black and Hispanic racial and ethnic groups, vaccine completion corresponded with a lower mean income-poverty ratio. For mixed ethnic groups, vaccine completion corresponded with a higher mean income-poverty ratio.

|  |
| --- |
| Figure 14. Income-poverty level versus ethnic group |

# 3. Analysis

We begin with various socioeconomic predictors that interest us. Some of these predictors are a subset of the other, or may be used to describe an overall category. For this reason

The variables used were chosen based on their interpretability and applicability to our research question “Which socioeconomic and geographic factors impact HPV vaccine completion rates among teenagers?” . There are seven potential outcome variables for HPV vaccine completion:

|  |
| --- |
| Table 3. 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 significantly high with a single dose. This recommendation was given with the goal of increasing vaccine coverage in areas with lower vaccine accessibility and availability (4). In September of 2023, PAHO followed suit in recommending a single-dose HPV vaccine regime for the Americas (5).

The predictor variables we chose to examine are listed below:

|  |
| --- |
| Table 4. 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.

To determine which predictor variables will be best suited for a general linear model, I will do a step-wise comparison and then measure the resulting models with AIC and WAIC.

# 4. References

1. Sciences, N. A. of, Engineering, & Medicine, and. (2018, March 1). Factors that affect health-care utilization. Health-Care Utilization as a Proxy in Disability Determination. https://www.ncbi.nlm.nih.gov/books/NBK500097/
2. 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
3. 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
4. https://www.who.int/news/item/20-12-2022-WHO-updates-recommendations-on-HPV-vaccination-schedule
5. https://www.paho.org/en/news/5-9-2023-paho-technical-advisory-group-recommends-countries-americas-use-single-dose-hpv